TILBURG UNIVERSITY

Tilburg University

Bayesian structural equation modeling

van Erp, S.J.

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
van Erp, S. J. (2020). *Bayesian structural equation modeling: The power of the prior*. Gildeprint.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Bayesian Structural Equation Modeling

The Power of the Prior

Sara van Erp

Original content © 2020 Sara van Erp, CC-BY 4.0
Chapter 2 © 2017 American Psychological Association
Chapter 4 © 2019 Elsevier Inc. All rights reserved.

The research in all chapters was funded by the Netherlands Organization for Scientific Research (NWO) through a Research Talent Grant (406-15-264).

Printed by: Gildeprint
ISBN: 9789464023978

Bayesian Structural Equation Modeling

The Power of the Prior

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. K. Sijtsma,
in het openbaar te verdedigen ten overstaan van een door het college voor
promoties aangewezen commissie in de Aula van de Universiteit op vrijdag 11
september 2020 om 13:30 uur
door Sarah Josephina van Erp, geboren te Breda.

Promotor:	Prof. Dr. J.K. Vermunt
Copromotores:	Dr. Ir. J. Mulder
	Dr. D.L. Oberski
Promotiecommissie:	Prof. Dr. Y. Rosseel
	Prof. Dr. A.G.J. van de Schoot
	Prof. Dr. M.C. Kaptein
	Dr. R.A. Kievit
	Dr. S. Jak

Contents

1 Introduction **4**

2 Prior sensitivity analysis in default Bayesian structural equation modeling **12**

2.1 Introduction 14
2.2 A structural equation model 17
2.3 Default priors for Bayesian SEM 19
2.4 A simulation study of default BSEM analyses 27
2.5 A practical guide to prior sensitivity analysis 44
2.6 Empirical application: democracy and industrialization data 49
2.7 Discussion 55

3 Bayesian multilevel structural equation modeling: An investigation into robust prior distributions **60**

3.1 Introduction 62
3.2 Empirical application 64
3.3 Bayesian doubly latent ordinal multilevel model 65
3.4 Robust prior distributions for random effects variances 68
3.5 Priors applied to the empirical example 73
3.6 Simulation studies 76
3.7 Discussion 104

4 Shrinkage priors for Bayesian penalized regression. **108**

4.1 Introduction 110
4.2 Bayesian penalized regression 113
4.3 Overview shrinkage priors 116
4.4 Illustrating the behavior of the shrinkage priors 126
4.5 Simulation study 132
4.6 Empirical applications 140
4.7 Discussion 148

5 A tutorial on Bayesian penalized regression with shrinkage priors for small sample sizes **150**
5.1 Introduction . . . 152
5.2 Running example: communities and crime . . . 155
5.3 Software . . . 155
5.4 Shrinkage priors . . . 156
5.5 Practical considerations . . . 160

6 Shrinkage priors for Bayesian measurement invariance: A robust approach for modeling and detecting non-invariance **166**
6.1 Introduction . . . 168
6.2 The Bayesian multiple group confirmatory factor model . . . 171
6.3 Prior distributions to model measurement invariance . . . 173
6.4 Illustration . . . 180
6.5 Application . . . 185
6.6 Discussion . . . 189

7 Epilogue **194**

References **200**

Summary **224**

Acknowledgements **228**

Chapter 1

Introduction

Structural equation modeling (SEM) is an important framework within the social sciences that encompasses a wide variety of statistical models. In its most simple form, path analysis allows for investigation of relations between observed or manifest variables, including mediation and/or moderation effects. Often, however, researchers are interested in unobserved or latent variables. Examples include personality, intelligence, depression, and values. Factor analysis provides a statistical technique to infer relationships between manifest and latent variables through specification of a measurement model. Additionally, a structural model can be specified in which causal relationships between multiple latent variables are hypothesized. Furthermore, a hierarchical or multilevel structure can be present when latent variables are investigated in multiple groups such as countries or classes, or across time. As a result, SEM provides an extremely powerful toolbox that enables researchers to translate many substantive theories into a statistical model and estimate the parameters of interest. Traditionally, estimation of SEMs has relied on maximum likelihood (ML; Jöreskog, 1969). Since ML maximizes the likelihood function, it results in estimates for which the observed data are most probable given the model. This approach works well in many cases. Unfortunately, there also exist a variety of situations in which ML performs subpar. Broadly, the problems with ML can be divided into three categories: 1) problems due to empirical underidentification; 2) sample size requirements; and 3) model flexibility.

First of all, empirical underidentification can lead to various problems due to unstable estimation (Rindskopf, 1984). Empirical underidentification occurs when the model is identified in theory, but not in practice given the data at hand (Kenny, 1979). For example, suppose we have a simple one-factor model with three indicators which is, given the usual restrictions, exactly identified in theory. However, if one of the loadings is close to zero in the application at hand, this value will be used in the denominator of the formulas to estimate the other two loadings and lead to unstable estimates with large sampling variance. Furthermore, the residual variances for those two indicators are influenced and can have large sampling variance and possibly negative estimates (i.e., Heywood cases). In extreme cases, these problems might lead to nonconvergence of the model. For example, Revilla and Saris (2013) investigated the occurrence of nonconvergence and Heywood cases in two rounds of the European Social Survey for split ballot-multitrait multimethod models and found that nonconvergence occurred in 30% of the cases while Heywood cases occurred in 46.7% of the cases. These problems can be traced back to rank deficiencies in the model (Oberski, 2019).

Second, ML and other classical estimation methods such as generalized least squares (GLS) or weighted least squares (WLS) rely on asymptotic theory and will therefore not necessarily result in valid inferences when used on small samples. For

large samples, the sample covariance matrix converges to the population covariance matrix. Only then will the statistical properties of the estimators hold, will the standard errors be estimated accurately, and will the distributions of test and fit statistics be known. For example, Hoogland and Boomsma (1998) conclude that, in general, a sample size of at least 500 observations is needed to obtain accurate standard errors. Furthermore, for the chi-square test, the sample size should be at least five times the degrees of freedom of the model to obtain correct type 1 error rates, in case of normal observed variables. Note that most of the studies included in Hoogland and Boomsma (1998) are based on relatively simple, single-level confirmatory factor models. In more complex multilevel SEMs the sample size requirements for ML can become even more impractical. For example, Meuleman and Billiet (2009) concluded that at least 60 groups are needed to detect large structural effects at the between level in a general multilevel SEM, whereas more than 100 groups are necessary to detect smaller structural effects. This resembles the recommendation by Hox and Maas (2001) who caution against the use of multilevel SEM with less than 100 groups. Aside from low power to detect effects, a small number of groups inadvertently influences the parameter estimates as well. Specifically, the between-level variance components are generally underestimated with standard errors that are too small (Hox & Maas, 2001).

The third problem with classical estimation of SEMs is that it limits the scope of models that can be considered. Certain SEMs that are realistic in practice are computationally inconvenient or impossible to estimate using ML. For example, a simple structure is generally assumed in confirmatory factor analysis such that each indicator loads on only one factor with zero so-called cross-loadings on the other factors. Assuming that all cross-loadings are exactly zero might be too restrictive, however, and it might be more realistic to allow indicators to have small loadings on factors other than the factor they are hypothesized to load on (B. O. Muthén & Asparouhov, 2012). Unfortunately, it is not possible to estimate all cross-loadings freely using ML since this leads to a nonidentified model. In a similar vein, a researcher might wish to estimate the correlations between the residuals of the factor indicators. This is the case in the classical SEM example in which the influence of industrialization in 1960 is measured on political democracy in 1960 and 1965 (Bollen, 1980). In this application, the indicators for political democracy are based on expert ratings, some of which come from the same expert in 1960 and 1965. For these ratings, we might expect the residuals to be correlated. Again, estimating all these correlations freely is not possible within the ML framework since it results in a nonidentified model (B. O. Muthén & Asparouhov, 2012). Apart from certain SEMs not being identified, models that rely on numerical integration further restrict the scope of feasible models in the traditional framework. This limitation is especially

relevant in multilevel SEM in which the random effects need to be numerically integrated out whenever the likelihood does not have a closed form expression and as the number of random effects increases, the dimension of numerical integration increases. As a result, computational methods traditionally associated with ML estimation will become slow and less precise, eventually resulting in convergence issues. For example, when estimating a multilevel SEM with the latent centering approach, ML will run into issues when the number of random effects exceeds four (Asparouhov & Muthén, 2019). Similarly, in the case of categorical data, ML is limited to four latent variables and no residual correlations between the categorical indicators or the analysis becomes computationally infeasible (Asparouhov & Muthén, 2007, 2010).

These limitations of ML have led researchers to turn to alternative estimation methods. In particular, Bayesian estimation of SEMs or BSEM has recently gained popularity (see, for example, Lee, 2007; B. O. Muthén & Asparouhov, 2012; Scheines, Hoijtink, & Boomsma, 1999). The Bayesian framework is one of the main approaches to statistical estimation and inference, dating back to the efforts of Thomas Bayes and Pierre-Simon Laplace in the eighteenth century (see, for a historical overview, Fienberg, 2006). It has not become popular, however, until more recent advances in computing that allowed for the use of Bayesian methods in non-trivial and more realistic applications. By this time, the well-established frequentist framework had become the primary approach to statistics. However, the use of Bayesian estimation has been rising steadily (Depaoli & van de Schoot, 2017, Figure 1).

There exist several important differences between the Bayesian and the frequentist frameworks. Here, three main differences will be briefly discussed. For more extensive introductions to Bayesian analysis the reader is referred to introductory textbooks on the topic, such as Gelman, Carlin, Stern, and Rubin (2004) or the annotated reading list by Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2017). The first difference is of a philosophical nature and lies in the interpretation of probability. In the frequentist framework, probability is defined as a long-run relative frequency of an event. Within the Bayesian framework, however, the notion of probability is used more broadly as a way to quantify uncertainty about the state of the world. The advantage of this more extensive view on probability is that it allows Bayesians to make probability statements about the parameters of a statistical model. Consequently, the results of a Bayesian analysis can be interpreted in a more intuitive manner. Specifically, given a frequentist 95% confidence interval around some parameter, we cannot conclude that there is 95% probability that the true population value lies within the interval. Instead, the interpretation of the frequentist confidence interval relies on the idea that if we would repeat the study multiple times then in 95% of the cases we would expect the resulting confidence

intervals to contain the true population value, but for a given confidence interval based on observed data we cannot make probabilistic statements in the frequentist domain. A Bayesian 95% credibility interval, on the other hand, does allow the intuitive interpretation of being the interval in which the true value lies with 95% probability.

The second difference arises in the methods used to estimate parameters. Frequentists estimate parameters by considering the value that is most likely given the data at hand (i.e., maximum likelihood estimation). However, parameters are ultimately viewed as fixed quantities and, due to the frequentist interpretation of probability cannot be assigned probability distributions. Bayesians can assign probability distributions to parameters and do so to represent the uncertainty about the parameters before observing the data. The resulting prior distribution ideally reflects the available information about the problem at hand and is subsequently updated by the data through Bayes' theorem, which results in the posterior distribution. Specifically, given data Y and a vector of model parameters $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|Y) = \frac{p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(Y)}, \tag{1.1}$$

here, $p(\boldsymbol{\theta}|Y)$ is the posterior distribution, $p(\boldsymbol{\theta})$ denotes the prior distribution, and $p(Y|\boldsymbol{\theta})$ denotes the likelihood of the data Y. The marginal likelihood $P(Y)$ is the distribution of the data marginalized over the parameters in $\boldsymbol{\theta}$, i.e., $\int_{\boldsymbol{\theta}} p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta})$, and serves as a normalizing constant.

The third main difference between the Bayesian and frequentist frameworks lies in the tools used for estimation. The ML method used in the frequentist framework relies on optimization to find the parameter estimates, i.e., the maximum of the likelihood. Furthermore, a local quadratic approximation is used to estimate the sampling variance which, although being exact in the limit for correct models, can lead to unreliable standard errors, confidence intervals, and p-values when the likelihood is not well-behaved. Bayesian estimation, on the other hand, generally uses Markov Chain Monte Carlo (MCMC) sampling to directly draw observations from the posterior distribution. The main reason for relying on MCMC sampling is the fact that computing the integral necessary for the marginal likelihood $P(Y)$ is difficult in any but the most simple, trivial models.

These differences lead to several advantages of the Bayesian framework in general and, more specifically, in the context of SEM. An advantage of the use of MCMC estimation is that it does not rely on approximations of the posterior distribution that might not be realistic in practice, but rather, it directly draws samples from the posterior distribution regardless of its form. Furthermore, the use

of MCMC estimation allows credibility intervals to be computed in a straightforward manner, even for functions of parameters. This is especially advantageous in SEM where we are often interested, for example, in indirect effects and the uncertainty around them. MCMC sampling obtains draws from the posterior distribution for each parameter. For an indirect effect, we can simply multiply the draws for the paths composing the effect to obtain the posterior distribution for the indirect effect. We can then compute any summary statistic of interest, such as the posterior mean, mode, median, standard deviation, or quantiles to obtain the credibility interval. Note that, unlike frequentist confidence intervals, this computation does not rely on normality assumptions but takes the true form of the posterior distribution into account (see e.g., Y. Yuan & MacKinnon, 2009). Finally, the main advantage of the Bayesian framework, especially in SEM, lies in the prior distribution.

As can be seen from Bayes' theorem in (1.1), the posterior distribution is a combination of the prior distribution and the likelihood of the data. As a result, the information in the posterior is a compromise between the information from these two sources. The relative contribution of the prior and the likelihood depends on the amount of information in each of these sources. Specifically, a larger sample provides more information and the corresponding likelihood will therefore have a stronger influence on the posterior compared to a smaller sample. Similarly, the prior distribution can vary in the amount of information it contains. A prior distribution that is more peaked contains more information compared to a prior distribution that is more flat or spread out.

In general, we can distinguish four types of prior distributions based on the type and amount of information they contain: subjective, objective, regularization, and data-dependent priors. Subjective priors contain information about the parameters based on previous research or expert knowledge and are therefore very application-specific. Although accurately specified subjective priors ultimately perform best (see e.g., Depaoli, 2012, 2013, 2014; Depaoli & Clifton, 2015), it is difficult to elicit subjective priors and specifically to accurately translate uncertainty in a probability distribution (e.g., Garthwaite, Kadane, & O'Hagan, 2005; Tversky, 1974). As a result, applied researchers often rely on so-called "objective" priors instead (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Objective priors (Berger, 2006) do not contain external subjective information and can therefore be used in an automatic fashion for a Bayesian data analysis. The third category, regularization or "weakly informative" priors, can be seen as a compromise between objective and subjective priors, since they include only a small amount of external information. For example, the prior can include the information that a factor loading generally does not exceed some value or the prior can incorporate a prior guess regarding the sparsity of the model. While this small amount of information

can help stabilize the estimation, regularization priors do not depend as heavily on the application at hand as subjective priors and are therefore more widely applicable. The final type of prior is the data-dependent or empirical Bayes prior. Like regularization priors, empirical Bayes priors do include external information, however, this information is based on the data at hand (see e.g., Carlin & Louis, 2000a, Chapter 3). Empirical Bayes priors can vary in their informativeness. Throughout this thesis, the main focus will be on regularization or "weakly informative" priors. However, we will also discuss the other types of priors, mainly in Chapter 2.

The prior distribution used in BSEM has the power to solve the issues of ML. First of all, since the prior is an additional source of information, it can ameliorate problems due to empirical underidentification (Can, van de Schoot, & Hox, 2014; Dagne, Howe, Brown, & Muthén, 2002; Kohli, Hughes, Wang, Zopluoglu, & Davison, 2015). Similarly, the capability of the prior to add information to the analysis reduces the minimum sample size requirements compared to ML (Depaoli & Clifton, 2015; Hox, van de Schoot, & Matthijsse, 2012). Finally, models that are not identified in a classical sense can be estimated in the Bayesian framework through careful specification of the prior distribution (Gustafson, 2010). These advantages can be accomplished through the specific type and amount of information that is incorporated in the prior distribution. However, it is currently unclear exactly how to specify the prior distribution in order to attain these advantages. Therefore, the goal of this thesis is to investigate prior specification in BSEM with a specific focus on how we can use the prior to advance the field of SEM.

The first part of this thesis focuses on prior distributions for variance parameters. It is well known in the Bayesian literature that variance parameters, especially at higher levels in the model, are highly sensitive to the prior distribution. Chapter 2 investigates the use of so-called "default" priors in a general SEM. Default priors are objective priors that can be used in an automatic fashion. We consider various default priors, as well as novel empirical Bayes priors. We find in Chapter 2 that the results from a BSEM analysis are generally quite sensitive to the specific default prior used, especially in small samples. Therefore, we provide guidelines on how to conduct a prior sensitivity analysis in BSEM to assess the sensitivity of the results to the prior distribution. In Chapter 3, we examine more robust alternatives to the default priors considered in Chapter 2. Traditionally, Bayesian analysis has relied on the use of conjugate prior distributions since the resulting posterior will have a known distributional form which eases computation. Nowadays, however, there is no longer a need to rely on conjugate priors due to the advance of sophisticated software packages. This opens up possibilities to use priors with other distributional forms that have more robust properties, such as heavier tails. In Chapter 3, these more robust priors are investigated specifically for variance parameters in a

multilevel SEM.

The second part of this thesis sets out to investigate the possibilities of using shrinkage priors in SEM. Shrinkage priors have specific characteristics that enable them to shrink small coefficients towards zero, while keeping large coefficients away from zero. By doing so, shrinkage priors have the ability to automatically perform variable selection. Many different shrinkage priors exist, which have mainly been applied in simple regression models. In Chapter 4, we review the literature on shrinkage priors and combine the most popular shrinkage priors in a comprehensive overview. Furthermore, we compare the performance of the shrinkage priors in terms of prediction accuracy and variable selection in a simple linear regression model. The aim of Chapter 4 is to better understand the behaviors and characteristics of shrinkage priors in a simple model, which can aid us in later applying the shrinkage priors in more complex SEMs. Chapter 5 can be seen as a practical translation of Chapter 4. In Chapter 5, the findings of Chapter 4 are discussed in a manner that is accessible for applied researchers, with a focus on software and application. Next, in Chapter 6, we apply two specific shrinkage priors, the spike-and-slab and the regularized horseshoe prior, to a multiple group confirmatory factor model. The goal of this chapter is to use the shrinkage prior to automatically model measurement invariance, which is an important concept whenever latent variables are compared across groups. This is a prime example of using the prior to identify a model that would not be identified in a classical sense and by doing so a more robust and flexible method is obtained to model measurement invariance. Finally, Chapter 7 concludes this thesis with a discussion of the work so far and a look to the future of BSEM.

Chapter 2

Prior sensitivity analysis in default Bayesian structural equation modeling

Based on van Erp, S., Mulder, J., and Oberski, D.L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363-388. doi:10.1037/met0000162

Abstract

Bayesian structural equation modeling (BSEM) has recently gained popularity because it enables researchers to fit complex models while solving some of the issues often encountered in classical maximum likelihood (ML) estimation, such as nonconvergence and inadmissible solutions. An important component of any Bayesian analysis is the prior distribution of the unknown model parameters. Often, researchers rely on default priors, which are constructed in an automatic fashion without requiring substantive prior information. However, the prior can have a serious influence on the estimation of the model parameters, which affects the mean squared error (MSE), bias, coverage rates, and quantiles of the estimates.

In this paper, we investigate the performance of three different default priors: noninformative improper priors, vague proper priors, and empirical Bayes priors, with the latter being novel in the BSEM literature. Based on a simulation study, we find that these three default BSEM methods may perform very differently, especially with small samples. A careful prior sensitivity analysis is therefore needed when performing a default BSEM analysis. For this purpose, we provide a practical step-by-step guide for practitioners to conducting a prior sensitivity analysis in default BSEM. Our recommendations are illustrated using a well-known case study from the structural equation modeling literature and all code for conducting the prior sensitivity analysis is made available in the online supplemental material.

Keywords: Bayesian, Structural Equation Models, Default Priors, Sensitivity Analysis.

2.1 Introduction

Psychologists and social scientists often ask complex questions regarding group- and individual differences and how these change over time. These complex questions necessitate complex methods such as structural equation modeling (SEM); its Bayesian version (Bayesian structural equation modeling; BSEM), in particular, has recently gained popularity (e.g., Kaplan, 2014) because it potentially resolves some of the difficulties with traditional frequentist SEM. For example, frequentist estimation of multilevel SEMs–often employed when studying multiple classrooms, schools, or countries–has been found to perform badly in terms of bias and power with a small number of groups (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Maas & Hox, 2005; Meuleman & Billiet, 2009; Ryu & West, 2009), while BSEM performed well even with small samples (Depaoli & Clifton, 2015; Hox et al., 2012). BSEM may also reduce issues with nonconvergence (Kohli et al., 2015) and inadmissible estimates (Can et al., 2014; Dagne et al., 2002), it is computationally convenient for models with many latent variables (Harring, Weiss, & Hsu, 2012; Lüdtke, Robitzsch, Kenny, & Trautwein, 2013; Oravecz, Tuerlinckx, & Vandekerckhove, 2011), and BSEM easily yields credible intervals (i.e., the Bayesian version of a confidence interval) on functions of parameters such as reliabilities (Geldhof, Preacher, & Zyphur, 2014) or indirect effects (Y. Yuan & MacKinnon, 2009). Furthermore, BSEM allows researchers to assume that traditionally restricted parameters, such as cross-loadings, direct effects, and error covariances, are approximately rather than exactly zero by incorporating prior information (MacCallum, Edwards, & Cai, 2012; B. O. Muthén & Asparouhov, 2012).

However, to take advantage of BSEM, one challenge must be overcome (MacCallum et al., 2012): the specification of the prior distributions. Prior specification is an important but difficult part of any Bayesian analysis. Ideally, the priors should accurately reflect preexisting knowledge about the world, both in terms of the facts and the uncertainty about those facts. Previous research has shown that BSEM has superior performance to frequentist SEM from a subjective Bayesian perspective when priors reflect researchers' beliefs exactly; and from a frequentist perspective, BSEM outperforms frequentist SEM when priors reflect reality. Priors that do not reflect prior beliefs to infinite accuracy (Bayesian perspective), or that do not correspond to reality (frequentist perspective), however, can lead to severe bias (Baldwin & Fellingham, 2013; Depaoli, 2012, 2013, 2014; Depaoli & Clifton, 2015). Moreover, eliciting priors is a time-consuming task, and even experts are often mistaken and prone to overstating their certainty (e.g., Garthwaite et al., 2005; Tversky, 1974). Additionally, in BSEM it is generally more difficult to specify subjective priors due to the many parameters, some of which are not easily interpretable (e.g., latent variable variances). Therefore, instead of relying fully on expert judgements, re-

searchers employing Bayesian analysis often use "default" priors. Default priors can be viewed as a class of priors that (i) do not contain any external substantive information, (ii) are completely dominated by the information in the data, and (iii) can be used in an automatic fashion for a Bayesian data analysis (Berger, 2006). For this reason default priors seem particularly useful for SEM because they allow us to use the flexible Bayesian approach without needing to translate prior knowledge into informative priors.

Previous research has investigated the performance of several default priors for BSEM. Thus far, the BSEM priors studied have been limited to proper priors chosen to equal the true population values in expectation, or chosen purposefully to be biased in expectation by a certain percentage (Depaoli, 2012, 2013, 2014; Depaoli & Clifton, 2015). These studies yielded important insights into the consequences of prior choice. However, commonly suggested alternative default priors in the Bayesian literature, such as noninformative improper priors and empirical Bayes priors (e.g., Carlin & Louis, 2000a; Casella, 1985; Natarajan & Kass, 2000), remain, to our knowledge, uninvestigated. Moreover, while several authors agree that any BSEM analysis should be accompanied by a sensitivity analysis, the available practical guidelines to do so focus on the situation in which substantive information was used to specify the prior (Depaoli & van de Schoot, 2017). In addition, Depaoli and van de Schoot (2017) provide specific guidelines on checking the sensitivity of the results to different inverse Wishart priors for the covariance matrix. Our contribution is that we specifically focus on prior specification in BSEM, we focus on default prior specification when prior information is weak or completely unavailable, and we specifically focus on univariate priors (i.e., univariate normal, and inverse gamma) which are easiest to interpret by applied researchers.

In addition to the lack of knowledge regarding priors in BSEM, there appears to be a lack of awareness of the importance of the prior as well. A recent review by van de Schoot et al. (2017) identified trends in and uses of Bayesian methods based on 1579 papers published in psychology between 1990 and 2015. Of the 167 empirical papers employing regression techniques (including SEM), only 45% provided information about the prior that was used. 31% of the papers did not discuss which priors were used at all, while 24% did not provide enough information to reconstruct the priors. In terms of the type of prior, 26.7% of the empirical papers used informative priors, of which only 4.5% (2 papers) employed empirical Bayes methods to choose the hyperparameters.

The literature review of van de Schoot et al. (2017) showed that a substantial part of Bayesian analyses in psychology relies on default priors. Now the problem is that the exact choice of the default prior may affect the conclusions substantially, as has been shown in the general Bayesian literature (e.g., Gelman, 2006; Lambert,

Sutton, Burton, Abrams, & Jones, 2005) and will be shown in the context of BSEM in this paper. Different software packages have implemented different default or weakly informative priors as their default software settings (Table 2.1). With the development of more user-friendly Bayesian software, more non-expert users are trying out Bayesian analysis in general and BSEM in particular and rely on the default software settings, without being fully aware of the influence and importance of the prior distributions. In the 167 empirical papers identified by van de Schoot et al. (2017), WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) was the most popular software program until 2012, but since 2013 this position has been taken over by the commercial SEM software Mplus (which has Bayesian methods implemented since 2010; L. K. Muthén & Muthén, 1998-2012; van de Schoot et al., 2017).

Table 2.1: Overview of default priors in software packages for BSEM

Software	Type of parameter	Default prior form	Default prior hyperparameters
Mplus (L. K. Muthén & Muthén, 1998-2012)	Intercepts/loadings/slopes	Normal	$N(0, 10^{10})$
	Thresholds	Normal	$N(0, 5)$
	Variance parameters	inverse Gamma	$IG(-1, 0)$
	Variance covariance matrices continuous variables	inverse Wishart	$IW(0, -p-1)$
	Variance covariance matrices categorical variables	inverse Wishart	$IW(I, p+1)$
	Class proportions mixture models	Dirichlet	$D(10, 10, \ldots, 10)$
Blavaan (Merkle & Rosseel, 2018)	Measurement intercepts	Normal	$N(0, 1000)$
	Structural intercepts/loadings/regression coefficients	Normal	$N(0, 100)$
	Precision residuals	Gamma	$G(1, 0.5)$
	Blocks of precision parameters	Wishart	$W(I, p+1)$
	Correlations	Beta*	$B(1, 1)$
Stan (Carpenter et al., 2017)	All parameters	Uniform	
Amos (Arbuckle, 2013)	All parameters	Uniform	
WinBUGS (Lunn et al., 2000)	All parameters	No defaults; manually specify proper priors	
JAGS (Plummer, 2003)	All parameters	No defaults; manually specify proper priors	

Note. p represents the number of variables. I represents the identity matrix. * The Beta prior for correlations in blavaan has support (-1, 1) instead of (0, 1).

This paper aims to further develop the practice and utility of default BSEM and to raise awareness that the exact choice of default prior in BSEM matters. Specifically, this paper has the following three goals. First, we propose two novel

empirical Bayes (EB) prior settings which adapt to the observed data and are easy to implement. Second, we investigate the performance of several default priors, including the novel EB priors, and compare them with the priors studied thus far, thereby investigating prior sensitivity in default BSEM. Third, since the choice of the default prior can have a large effect on the estimates in small samples, we provide a step-by-step guide on how to perform a default prior sensitivity analysis. Note that we focus on frequentist properties of the different default priors, such as bias, mean squared errors, and coverage rates. We take this perspective because it is common to focus on frequentist properties when assessing the performance of default priors (Bayarri & Berger, 2004). A different and popular perspective on Bayesian statistics is that of updating one's prior beliefs with the data. In this perspective, instead of default priors, informative priors are used which contain external subjective information about the magnitude of the parameters before observing the data. Specification of such subjective priors will not be explored in the current paper.

The rest of this article is organized as follows. We first introduce the BSEM model using a running example from the SEM literature. In the subsequent section, we discuss possible priors that have been suggested both in the BSEM and in the wider Bayesian analysis literature. Subsequently, a simulation study investigates the effect these prior choices have on BSEM estimates. We then provide practical guidelines based on the results of the simulation for practitioners who wish to perform their own sensitivity analysis. Finally, we apply these guidelines to empirical data from the running example, providing a demonstration of sensitivity analysis in BSEM.

2.2 A structural equation model

Throughout this paper we will consider a linear structural equation model with latent variables from the literature. We have selected this model because it is one of the most popular example models in structural equation modeling. Furthermore the model includes a mediation effect, which is of interest in substantive research. As a result, investigation of this model will not only result in general insights regarding default priors in BSEM but will also provide specific information about default priors for mediation analysis. The model (Figure 2.1) describes the influence of the level of industrialization in 1960 (ξ) on the level of political democracy in 1960 (η^{60}) and 1965 (η^{65}) in 75 countries. Industrialization is measured by three indicators and the level of democracy by four indicators at each time point. The indicators for level of democracy consist of expert ratings, and, since some of the ratings come from the same expert at both time points or the same source in the same year, several measurement errors correlate, which we model through pseudo-latent variables $\boldsymbol{D}$,

following Dunson, Palomo, and Bollen (2005).

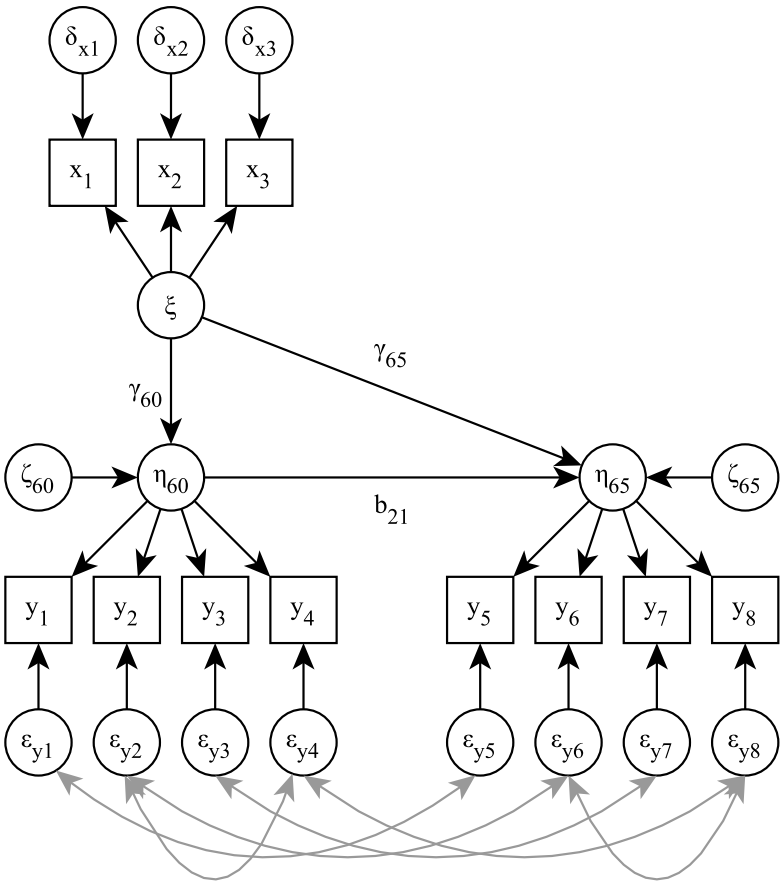

Figure 2.1: Structural equation model describing the influence of industrialization in 1960 (ξ) on political democracy in 1960 (η^{60}) and 1965 (η^{65}).

The structural model (for $i = 1, \ldots, n$) is given by:

$$\begin{aligned} \boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\xi_i + \boldsymbol{\zeta}_i \quad & \text{with } \xi_i \sim N(\mu_\xi, \omega_\xi^2), \\ & \text{and } \boldsymbol{\zeta}_i \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\zeta}}) \end{aligned}$$

The measurement model is given by:

$$\begin{aligned} \boldsymbol{y}_i = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + \boldsymbol{D}_i + \boldsymbol{\epsilon}_i^y \quad & \text{with } \boldsymbol{D}_i \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{D}}), \\ & \text{and } \boldsymbol{\epsilon}_i^y \sim N(\mathbf{0}, \boldsymbol{\Sigma}_y) \\ \boldsymbol{x}_i = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\xi_i + \boldsymbol{\delta}_i^x \quad & \text{with } \boldsymbol{\delta}_i^x \sim N(\mathbf{0}, \boldsymbol{\Sigma}_x) \end{aligned}$$

Here, the structural mean and intercepts μ_ξ and $\boldsymbol{\alpha}$ reflect the mean structure in the structural part of the model, while the measurement intercepts $\boldsymbol{\nu}_y$ and $\boldsymbol{\nu}_x$ reflect the mean structure in the measurement part of the model. The loadings $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ represent the relations between the latent variables and their indicators,

and the structural regression coefficients $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ represent the relations between the latent variables. The residual variances $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_x$ reflect the variation in the measurement errors, and the random variances ω_ξ^2, $\boldsymbol{\Omega_\zeta}$, and $\boldsymbol{\Omega_D}$ reflect the variation in the latent variables. In practice, researchers are often most interested in the relations between the latent variables, in this case the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$ of industrialization in 1960 on political democracy in 1965. Note that this specification is more restrictive than necessary since, following Dunson, Palomo, and Bollen (2005), we fix the nonzero entries in the weight matrix for $\boldsymbol{D}$ to be 1. As a result negative covariances are not allowed. A solution to this problem is implemented in the R-package blavaan (Merkle & Rosseel, 2018; R Core Team, 2015). The model is identified by restricting the first intercept and loading of each latent variable, i.e., $\lambda_y^1 = \lambda_y^5 = \lambda_x^1 = 1$ and $\nu_y^1 = \nu_y^5 = \nu_x^1 = 0$. The Appendix provides the full model in matrix form and a more detailed description of the data and the model can be found in Bollen (1980, 1989).

In the application we will use the original data containing observations from 75 countries (available in the lavaan package in R; Rosseel, 2012), and a subset of the data containing only the first 35 observations. Maximum likelihood (ML) estimation for this subset gives two warnings: 1) the standard errors of the parameter estimates may not be reliable due to a non-positive definite first-order derivative product matrix, and 2) the latent variable covariance matrix is not positive definite. The first warning can be an indication of weak empirical identification due to the small sample size, whereas the second warning indicates an inadmissible parameter estimate; in this case the estimated variance of the pseudo-latent variable representing the relation between ϵ_4^y and ϵ_8^y, i.e., $\hat{\omega}_{D48}^2$, is negative. These warnings clearly illustrate that when using classical ML estimation, researchers may encounter certain problems which may be overcome by adopting a Bayesian approach, where prior distributions are specified in the subspace of admissible solutions.

2.3 Default priors for Bayesian SEM

Default priors have the following three key properties. First, they do not contain external subjective information. Second, they are completely dominated by the data. Third, they can be used in an automatic fashion for a Bayesian data analysis. Because of the second property these priors are also referred to as "non-informative" or "weakly informative" priors. Such priors are often used in Bayesian analysis, including BSEM, when no substantive information is available or when the researcher does not wish to incorporate any substantive information through the prior distribution. Default priors allow researchers to use the powerful and flexible Bayesian approach without needing to specify an informative prior based on

one's prior knowledge, which can be a difficult and time-consuming task. Different software packages use different default priors in their automatic settings based on various heuristic arguments (see Table 2.1). For example, the commercial SEM software Mplus (L. K. Muthén & Muthén, 1998-2012) specifies a uniform improper prior for variance parameters by default, while the Bayesian modeling software WinBUGS (Lunn et al., 2000) recommends vague proper inverse Gamma priors for the variances.

Default priors can roughly be divided in the following three categories: noninformative[1] improper priors, vague proper priors, and empirical Bayes priors. The first two have been used extensively in the BSEM literature, while the latter has, to our knowledge, not been applied to BSEM yet, but it is popular in the general literature on Bayesian modeling (Carlin & Louis, 2000a; Casella, 1985; Natarajan & Kass, 2000). In this study, we will focus on different priors from each of these three commonly used types of default priors. Some priors are chosen because they are the default setting in popular software programs (specifically Mplus and blavaan), while other priors are chosen because they are widely used, or because they have been shown to perform well in certain situations. The EB priors are novel in BSEM and have been included to investigate whether an EB approach can be advantageous in BSEM.

For all three types, we focus on priors that have a conditionally conjugate form. Conditionally conjugate priors have the advantage that they result in fast computation because the resulting conditional posteriors have known distributional forms (i.e., they have the same distribution as the prior) from which we can easily sample. Specifically, the conditionally conjugate prior for a location parameter (e.g., intercepts, loadings, and regression coefficients) is the normal distribution, and for a variance parameter it is the inverse Gamma distribution. Note that these are univariate priors. Generally, it is possible to instead specify a multivariate normal distribution for the vector of location parameters and an inverse Wishart prior for the covariance matrix. However, in this paper we focus specifically on univariate priors for separate parameters. We will now discuss these different default priors in more detail.

Noninformative improper priors

Noninformative improper priors are most commonly used in "objective Bayesian analysis" (Berger, 2006). In a simple normal distribution with unknown mean μ and unknown variance σ^2, for example, the standard noninformative improper prior

[1] Here, noninformative refers to the ultimate goal of the prior rather than its actual behavior. In fact, many different noninformative priors exists (Kass & Wasserman, 1996), which often result in slightly different estimates.

$p(\mu, \sigma^2) \propto \sigma^{-2}$ (known as Jeffreys' prior) yields exactly the same point and interval estimates for the population mean as does classical ML estimation; hence the name "objective Bayes". An improper prior is not a formal probability distribution because it does not integrate to unity. A potential problem of noninformative improper priors is that the resulting posteriors may also be improper, which occurs when there is too little information in the data (Hobert & Casella, 1996). In the above example of a normal distribution with unknown mean and variance we need at least two distinct observations in order to obtain a proper posterior for μ and σ^2 when starting with the improper Jeffreys' prior. Currently, little is known about the performance of these types of priors in BSEM. Throughout this paper we will therefore consider the following noninformative improper priors for variance parameters σ^2:

- $p(\sigma^2) \propto \sigma^{-2}$. This prior is most commonly used in objective Bayesian analysis for variance components. It is equivalent to a uniform prior on $\log(\sigma^2)$. There have been reports, however, that this prior results in improper posteriors for variances of random effects in multilevel analysis (e.g., Gelman, 2006). In a simple normal model with known mean and unknown variance, at least one observation is needed for this prior to result in a proper posterior for the variance.

- $p(\sigma^2) \propto \sigma^{-1}$. This prior was recommended by Berger (2006) and Berger and Strawderman (1996) for variance components in multilevel models. For this prior, at least two observations are needed in a normal model with known mean and unknown variance to obtain a proper posterior.

- $p(\sigma^2) \propto 1$. This prior is the default choice in Mplus (L. K. Muthén & Muthén, 1998-2012). Gelman (2006) noted that it may result in overestimation of the variance. When using this prior in a normal model with known mean and unknown variance, at least three observations are needed to obtain a proper posterior for the variance.

Each of these noninformative improper priors can be written as the conjugate inverse Gamma prior. The inverse Gamma distribution is given by:

$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \text{ with shape } \alpha > 0 \text{ and scale } \beta > 0$$

When the shape parameter $\alpha = 0$ and the scale parameter $\beta = 0$, we obtain $p(\sigma^2) \propto \sigma^{-2}$. When the shape parameter $\alpha = -\frac{1}{2}$ and the scale parameter $\beta = 0$, we obtain $p(\sigma^2) \propto \sigma^{-1}$. When the shape parameter $\alpha = -1$ and the scale parameter $\beta = 0$, we obtain $p(\sigma^2) \propto 1$.

Table 2.2 presents these priors for all variance components in our model. For the intercepts, means, loadings, and regression coefficients, the standard noninformative improper prior is the uniform prior from $-\infty$ to $+\infty$. The vague proper prior $N(0, 10^{10})$ approximates this uniform prior. Thus, for the intercepts, means, loadings and regression coefficients, we will only investigate vague proper and empirical Bayes priors, which are discussed next.

Table 2.2: Overview of the default prior specifications per parameter considered throughout this paper.

		Prior			
Parameter type	Parameter	Default	Noninformative improper	Vague proper	Empirical Bayes (EB)
Latent variable variances	ω_ξ^2	$\pi(\omega_\xi^2) \propto 1$	$\pi(\omega_\xi^2) \propto 1$	IG$(0.001, 0.001)$	IG $\left(\frac{1}{2}, \hat{\omega}_\xi^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$
			$\pi(\omega_\xi^2) \propto \omega_\xi^{-1}$	IG$(0.01, 0.01)$	
			$\pi(\omega_\xi^2) \propto \omega_\xi^{-2}$	IG$(0.1, 0.1)$	
	ω_ζ^2	$\pi(\omega_\zeta^2) \propto 1$	$\pi(\omega_\zeta^2) \propto 1$	IG$(0.001, 0.001)$	IG $\left(\frac{1}{2}, \hat{\omega}_\zeta^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$
			$\pi(\omega_\zeta^2) \propto \omega_\zeta^{-1}$	IG$(0.01, 0.01)$	
			$\pi(\omega_\zeta^2) \propto \omega_\zeta^{-2}$	IG$(0.1, 0.1)$	
	ω_D^2	$\pi(\omega_D^2) \propto 1$	$\pi(\omega_D^2) \propto 1$	IG$(0.001, 0.001)$	IG $\left(\frac{1}{2}, \hat{\omega}_D^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$
			$\pi(\omega_D^2) \propto \omega_D^{-1}$	IG$(0.01, 0.01)$	
			$\pi(\omega_D^2) \propto \omega_D^{-2}$	IG$(0.1, 0.1)$	
Residual variances	σ_y^2	$\pi(\sigma_y^2) \propto 1$	$\pi(\sigma_y^2) \propto \sigma_y^{-2}$	IG$(0.001, 0.001)$	IG $\left(\frac{1}{2}, \hat{\sigma}_y^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$
			$\pi(\sigma_y^2) \propto \sigma_y^{-1}$	IG$(0.01, 0.01)$	
			$\pi(\sigma_y^2) \propto \sigma_y^{-2}$	IG$(0.1, 0.1)$	
	σ_x^2	$\pi(\sigma_x^2) \propto 1$	$\pi(\sigma_x^2) \propto \sigma_x^{-2}$	IG$(0.001, 0.001)$	IG $\left(\frac{1}{2}, \hat{\sigma}_x^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$
			$\pi(\sigma_x^2) \propto \sigma_x^{-1}$	IG$(0.01, 0.01)$	
			$\pi(\sigma_x^2) \propto \sigma_x^{-2}$	IG$(0.1, 0.1)$	
Structural intercepts	α	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\alpha}^2 + \hat{\omega}_\zeta^2)$
			-	N$(0, 100)$	
Structural regression coefficients	b	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{b}^2 + \hat{\omega}_\zeta^2)$
			-	N$(0, 100)$	
	γ	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\gamma}^2 + \hat{\omega}_\zeta^2)$
			-	N$(0, 100)$	
Latent variable mean	μ_ξ	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\mu}_\xi^2 + \hat{\omega}_\xi^2)$
			-	N$(0, 100)$	
Measurement intercepts	ν_y	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\nu}_y^2 + \hat{\sigma}_y^2)$
			-	N$(0, 1000)$	
	ν_x	N$(0, 10^{10})$	-	N$(0, 10^{10})$	$N(0, \hat{\nu}_x^2 + \hat{\sigma}_x^2)$
			-	N$(0, 1000)$	
Loadings	λ_y	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\lambda}_y^2 + \hat{\sigma}_y^2)$
			-	N$(0, 100)$	
	λ_x	N$(0, 10^{10})$	-	N$(0, 10^{10})$	N$(0, \hat{\lambda}_x^2 + \hat{\sigma}_x^2)$
			-	N$(0, 100)$	

Q^{-1}: denotes the regularized inverse Gamma function.

Vague proper priors

A common solution to avoid improper posteriors while keeping the idea of noninformativeness in the prior is to specify vague proper priors. These priors are formal probability distributions, where the hyperparameters are chosen such that the information in the prior is minimal. In the case of variance parameters, vague proper priors can be specified as conjugate inverse Gamma priors with hyperparameters close to zero, typically 0.1, 0.01, or 0.001. These priors approximate the improper prior $p(\sigma^2) \propto \sigma^{-2}$ (Berger, 2006). The latter option, $IG(\epsilon, \epsilon)$, with $\epsilon = 0.001$ is used as example throughout the WinBUGS manual. We will consider these three typical prior specifications for the variance parameters in our model. Note that smaller hyperparameters lead to a prior that is more peaked around zero. For means and regression parameters, we will investigate a normal prior with a large variance. This vague proper prior approximates a flat prior. Specifically, we shall use the normal prior $N(0, 10^{10})$, which is the default in Mplus. In addition, we will consider the blavaan default setting for location parameters, which is the normal prior $N(0, 1000)$ for the measurement intercepts and the normal prior $N(0, 100)$ for the loadings, structural intercepts, and structural regression coefficients. We will label this prior coupled with $\pi(\sigma^2) \propto 1$ for variances "vague normal". The vague proper priors that will be considered throughout this paper are summarized in Table 2.2.

A potential problem of vague proper priors is that the exact hyperparameters are arbitrarily chosen, while this choice can greatly affect the final estimates. For example, there is no clear rule stating how to specify the shape and scale parameter of the inverse Gamma prior, namely, 0.1, 0.01, 0.001, or perhaps even smaller. Gelman (2006) showed that in a multilevel model with 8 schools on the second level, the posterior for the between-school variance was completely dominated by the inverse Gamma prior with small hyperparameters. In addition, the inverse Gamma prior depends on the scale of the data. Specifically, an $IG(0.1, 0.1)$ prior might be a noninformative choice when the data are standardized, but can be very informative when the data are unstandardized. It is yet unclear how this prior performs in structural equation models, which are considerably more complex than the 8 schools example studied by Gelman (2006), in terms of the number of parameters and the relations between them.

Empirical Bayes priors

The third type of default priors we consider is empirical Bayes (EB) priors. The central idea behind the EB methodology is that the hyperparameters are chosen based on the data at hand (see e.g., Carlin & Louis, 2000a, , Ch. 3). This results in a prior with substantial probability mass in the region where the likelihood is

concentrated. EB methodology can be seen as a compromise between classical and Bayesian approaches (Casella, 1992). Since all the data are used to inform the prior distribution, EB methods are useful for combining evidence (e.g., across neighborhoods, Carter and Rolph (1974); or across law schools, Rubin (1980)). In our application, we expect an EB prior informed by the data of all countries to provide better estimates because it adds more information to the analysis than improper or vague priors, or ML estimation. There is also a computational advantage of EB priors as noted by Carlin and Louis (2000b) who state that the Markov Chain Monte Carlo (MCMC) sampler based on EB priors can be more stable.

We focus on the parametric EB approach, in which a specific distributional form of the prior is assumed, typically conjugate, with only the hyperparameters unknown. Different methods have been proposed to obtain the hyperparameters in this setting. First the hyperparameters can be estimated using the marginal distribution of the data (i.e., the product of the likelihood and the prior integrated over the model parameters: $p(Data) = \int f(Data|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the vector with parameters, with prior $p(\boldsymbol{\theta})$, and f denotes the likelihood of the data given the unknown model parameters) (see e.g., Carlin & Louis, 2000a; Casella, 1985; Laird & Ware, 1982). An example is the EB version of the well-known g-prior of Zellner (1986). The g-prior is centered around a reference (or null) value where g controls the prior variance. In EB methodology, g is estimated from the marginal distribution (see e.g., Liang, Paulo, Molina, Clyde, & Berger, 2008, , and the references therein). The resulting EB prior can be seen as the best predictor of the observed data. The difficulty with this approach, however, is that an analytic expression of the marginal distribution of the data may not be available for large complex models. This is also the case in structural equation models, and therefore we will not use the marginal distribution of the data to construct an EB prior in this paper.

A second EB approach, which is simpler to implement, is to specify weakly informative priors centered around the estimates (e.g., ML) of the model parameters. These priors contain minimal information so that the problem of double use of the data is negligible. This type of EB prior has been investigated in generalized multilevel modeling (Kass & Natarajan, 2006; Natarajan & Kass, 2000) and structural equation modeling (Dunson et al., 2005). This approach, however, may perform badly when the estimates that are used for centering the EB priors are unstable, which is the case in structural equation modeling with small samples.

For this reason an alternative EB prior is proposed which is novel in the BSEM literature. The idea is to first center the prior around a reference value to minimize its dependence on unstable estimates. Subsequently, the other hyperparameters are estimated from the data from a simplified model to keep the solution tractable. The idea of this EB prior was inspired by the constrained posterior prior approach

of Mulder, Hoijtink, and Klugkist (2010); Mulder et al. (2009) for Bayesian model selection. As will be shown the proposed EB prior generally contains less information than the information of one observation. For this reason, the double use of the data and the resulting underestimation of the posterior variance, a known problem of EB methodology (e.g., Carlin & Louis, 2000a; Darnieder, 2011; Efron, 1996), is expected to be negligible.

EB priors for intercepts, means, factor loadings, and regression coefficients

Here we discuss the EB prior for the intercepts, means, factor loadings, and regression coefficients for which the conditionally conjugate prior is normally distributed. To estimate the hyperparameters, i.e., prior mean and the prior variance, we simplify the endeavor by constructing a normal prior for α, denoted by $N(\mu_\alpha, \tau_\alpha^2)$, for the following model

$$y_i = \alpha + \epsilon_i \text{ with } \epsilon_i \sim N(0, \sigma^2), \tag{2.1}$$

for $i = 1, \ldots, n$ observations, with unknown error variance σ^2. Note that α in the model in (1) denotes a location parameter, i.e., an intercept, mean, factor loading, or regression coefficient in the model considered in this paper. The prior mean μ_α is set equal to a reference (or null) point value to avoid the heavy dependence on the data. In the current setting, we set the prior mean equal to zero because of its special meaning of "no effect". The prior variance τ_α^2 is then estimated as the variance in a restricted model where the prior mean, $\mu_\alpha = 0$, is plugged in for α, i.e., $y_i \sim N(0, \tau_\alpha^2)$, for $i = 1, \ldots, n$. The variance in this model can be estimated as $\hat{\tau}_\alpha^2 = n^{-1} \sum_{i=1}^n y_i^2 \approx E[Y^2]$. Subsequently, an expression for the estimate $\hat{\tau}_\alpha^2$ can be obtained by deriving $E[Y^2]$ from the original model (2.1) according to

$$E[Y^2] = \text{ Var}(Y) + (E[Y])^2 = \sigma^2 + \alpha^2. \tag{2.2}$$

Thus, the prior variance is chosen to be equal to the sum of the estimated error variance and the square of the estimated effect, i.e., to $\hat{\tau}_\alpha^2 = \hat{\sigma}^2 + \hat{\alpha}^2$.

This prior has two important properties. First, this prior has clear positive support where the likelihood is concentrated, which is a key property of an EB prior. This can be seen from the fact that the prior variance will be large (small) when the difference between the observed effect and the prior mean, $\hat{\alpha}$, is large (small). Note however that by centering the prior around a reference value instead of the observed effect, the prior will be less sensitive to the instability of the ML estimates. Second, this EB prior does not contain more information than the information of a single observation. To see this note that the standard error of α is equal to $\sigma/\sqrt{n}$.

Furthermore, the EB prior standard deviation is smallest when $\hat{\alpha} = 0$, in which case $\hat{\tau}_\alpha = \hat{\sigma}$, which corresponds to the standard error based on a single observation. If $\hat{\alpha} \neq 0$, then $\hat{\tau}_\alpha > \hat{\sigma}$, which implies less information than the information in a single observation. For this reason, we expect the problem of using the data twice (i.e., for prior specification and estimation) to be negligible. This behavior is illustrated in Figure 2.2. In the case of no effect, i.e., $\hat{\alpha} = 0$ (Data 1), the prior variance is equal to the error variance. When $\hat{\alpha} \neq 0$ (Data 2), the prior variance becomes larger, i.e., the error variance plus the squared estimated effect, to ensure positive support around $\hat{\alpha}$. Note that the EB prior behaves similarly to the constrained posterior prior (Mulder et al., 2010) with the difference that the EB prior is simpler to compute.

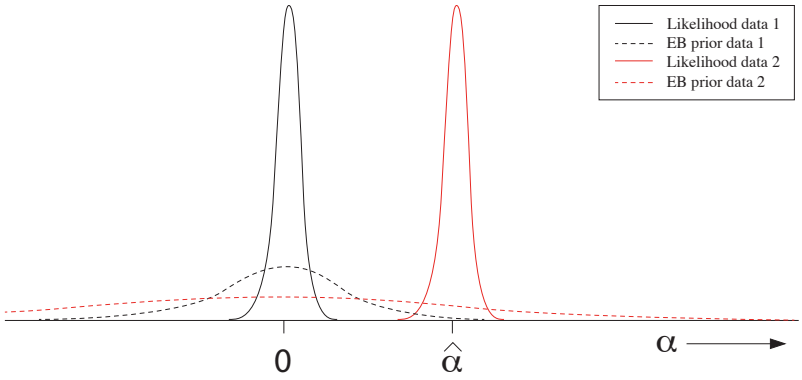

Figure 2.2: Illustration EB priors for location parameters. The EB prior for α is $N(0, \tau_\alpha^2)$ with τ_α^2 equal to $\hat{\sigma}^2$ for Data 1, and equal to $\hat{\sigma}^2 + \hat{\alpha}^2$ for Data 2.

This methodology will be used to construct EB priors for the intercepts, means, factor loadings, and regression coefficients. For example for the measurement intercept of y_2, denoted by ν_2^y, the EB prior variance is equal to the squared ML estimate, i.e., $(\hat{\nu}_2^y)^2$, plus the ML estimate of the variance of the error δ_2^y, i.e., $\hat{\sigma}_{y2}^2$, and thus, the EB prior is distributed as, $\nu_2^y \sim N(0, (\hat{\nu}_2^y)^2 + \hat{\sigma}_{y2}^2)$. An overview of all EB priors for location parameters is given in Table 2.2.

EB priors for variance components

The above methodology to construct EB priors for location parameters cannot be used for variance components because a clear reference (or null) value is generally unavailable for these types of parameters. Therefore we will consider two alternative approaches for the priors for the variances which will be combined with the EB prior for intercepts, means, factor loadings, and regression coefficients. In the first combination, which we will label EB1, the priors for the variance components are centered around the ML estimate, as was also considered by Natarajan and Kass (2000). Again we consider the conditionally conjugate prior which has an inverse

gamma distribution with a shape parameter and a scale parameter. The shape parameter controls the amount of prior information. By setting the shape parameter to $\frac{1}{2}$, the prior carries the information that is equivalent to one data point (Gelman et al., 2004, , p. 50). For this reason the double use of the data is also not a serious concern in this case. The scale parameter is chosen such that the prior median equals the ML estimate, i.e., $\hat{\sigma}^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})$ with $\hat{\sigma}^2$ denoting the ML estimate of the variance parameter and Q^{-1} denoting the regularized inverse Gamma function (see Table 2.2). A potential issue of this prior is it heavy dependence on the ML estimates of the variances. Therefore in the second combination, labeled EB2, noninformative uniform priors are specified for the variances. Thus, the EB2 prior can be seen as a hybrid EB prior, where the priors for location parameters depend on the data and the priors for variance parameters are completely independent of the data. [2]

2.4 A simulation study of default BSEM analyses

Even though all the discussed priors reflect some form of default noninformative BSEM analysis, each choice may result in different conclusions. A simulation study was set up to investigate the performance of the different default priors in the industrialization and political democracy model, a classical SEM application. A common method to check the performance of objective priors is to look at their frequentist properties (Bayarri & Berger, 2004). In particular, we were interested in (1) convergence of the Bayesian estimation procedure, (2) (relative) bias, (3) mean squared error (MSE), (4) coverage rates, (5) quantiles, and (6) type 1 error rates of the direct and indirect effects. We will end the section with a general conclusion regarding the performance of the different default priors.

For the data generation, we considered four different sample sizes: 35, 75, 150, and 500. For such a complex SEM, a sample size of 35 or 75 might seem extremely small. However, BSEM is often recommended especially in situations where the sample size is small (Heerwegh, 2014; Hox et al., 2012; Lee & Song, 2004). Furthermore, the original sample size was only 75. We expect the influence of the prior to decline as sample size increases. The population values were equal to the ML estimates of the original data (see the online supplemental material for an overview). We manipulated the population values for the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$, since these are the parameters of substantive interest in

[2] We also considered a third combination: an EB prior for the variances (i.e., an inverse Gamma prior centered around the ML estimate) combined with the $N(0, 10^{10})$ prior for the location parameters. Thus, this combination can also be seen as a hybrid EB prior, where the priors for the variance parameters depend on the data and the priors for the location parameters are completely independent of the data. However, this combination performed similarly to the uniform prior for variances combined with the $N(0, 10^{10})$ prior for location parameters and we have therefore not considered it further.

the model. We also manipulated two loadings of $\boldsymbol{y}$ (λ_4^y and λ_8^y) and the variances of the pseudo-latent variables $\boldsymbol{\Omega_D}$, which represent the measurement error correlations. These variances were manipulated because previous research indicates that the vague proper priors for variances are especially influential when the variance parameter is estimated to be close to zero (Gelman, 2006). The manipulations of the population values are shown in Table 2.3 and were selected in such a way that we obtained a wide range of values. We used a fractional design in which we simulated data under two combinations of population values: 1) combinations of population values for the direct and indirect effect ($3 \times 3 = 9$ conditions), and 2) combinations of population values for the loadings and variances of error correlations ($3 \times 2 = 6$ conditions). In total, this resulted in 15 different populations. From each population, we generated 500 datasets per sample size. [3] Table 2.3 presents an overview of all 60 data-generating conditions.

[3] We created cumulative average plots to assess whether 500 replications were enough to attain Monte Carlo convergence, which was the case.

Table 2.3: Overview of the data generating and analysis conditions included in the simulation study.

Variable	# Levels	Values	
Data generating conditions			
Sample size	4	$N \in \{35, 75, 150, 500\}$	
Direct effect	3	$\gamma_{65} = 0$	
		$\gamma_{65} = 1$	
		$\gamma_{65} = 2$	
Indirect effect	3	$\gamma_{60} \cdot b_{21} = 0 \times 0.837$	
		$\gamma_{60} \cdot b_{21} = 1 \times 0.837$	
		$\gamma_{60} \cdot b_{21} = 2 \times 0.837$	
Loadings	3	λ_4^y, $\lambda_8^y = 0$	
		λ_4^y, $\lambda_8^y = 1$	
		λ_4^y, $\lambda_8^y = 2$	
Error covariances	2	$\boldsymbol{\Omega_D = 0}$	
		$\boldsymbol{\Omega_D = 1}$	
Analysis conditions			
Priors	9	Noninformative improper:	
			$\pi(\sigma^2) \propto 1$ and $\text{N}(0, 10^{10})$ (Mplus default)
			$\pi(\sigma^2) \propto \sigma^{-1}$ and $\text{N}(0, 10^{10})$
			$\pi(\sigma^2) \propto \sigma^{-2}$ and $\text{N}(0, 10^{10})$
		Vague proper:	
			$\text{IG}(0.001, 0.001)$ and $\text{N}(0, 10^{10})$
			$\text{IG}(0.01, 0.01)$ and $\text{N}(0, 10^{10})$
			$\text{IG}(0.1, 0.1)$ and $\text{N}(0, 10^{10})$
		Vague normal:	
			$\text{N}(0, 1000)$ and $\text{N}(0, 100)$ and $\pi(\sigma^2) \propto 1$
		Empirical Bayes (EB):	
		EB1:	$\text{IG}\left(\frac{1}{2}, \hat{\sigma}^2 \cdot Q^{-1}(\frac{1}{2}, \frac{1}{2})\right)$ and $\text{N}(0, \hat{\mu}^2 + \hat{\sigma}^2)$
		EB2:	$\pi(\sigma^2) \propto 1$ and $\text{N}(0, \hat{\mu}^2 + \hat{\sigma}^2)$
Maximum Likelihood	1		

Note. The maximum likelihood (ML) estimates from the original data were: $\hat{\gamma}_{65} = 0.572$; $\hat{\gamma}_{60} \cdot \hat{b}_{21} = 1.483 \cdot 0.837$; $\hat{\lambda}_4^y = 1.265$; $\hat{\lambda}_8^y = 1.266$; $\hat{\omega}^2_{D_{15}} = 0.624$; $\hat{\omega}^2_{D_{24}} = 1.313$; $\hat{\omega}^2_{D_{26}} = 2.153$; $\hat{\omega}^2_{D_{37}} = 0.795$; $\hat{\omega}^2_{D_{48}} = 0.348$; and $\hat{\omega}^2_{D_{68}} = 1.356$.

Each condition was analyzed using the nine different default prior combinations with the same type of prior being specified for all parameters in the model at once, i.e., three noninformative improper priors, three vague proper priors, two EB priors, and the vague normal setting. Note that in the case of the noninformative improper and vague proper priors only the priors on the variance parameters change, while the priors on the mean and regression parameters are specified as the normal prior $\text{N}(0, 10^{10})$. In addition, ML estimation was included for each condition, leading to a total of $10 \times 15 \times 4 = 600$ conditions, as shown in Table 2.3.

The EB priors are based on the ML estimates, which sometimes included Heywood cases (i.e., an estimated negative variance). In the case of negative ML estimates for the variance parameters, we set the prior median for the EB prior on variance parameters equal to 0.001. Preliminary analyses showed that the precise choice had little effect on the posterior. For the mean, intercept, loading, and regression parameters the residual variances of the model equations were sometimes estimated to be negative, in which case we fixed them to zero for computation of the prior variance. Again, preliminary analyses indicated that the exact choice did not have any clear influence on the posterior, as long as the estimate was fixed to a small number, e.g., 0.001.

For each analysis we ran two MCMC chains and discarded the first half of each chain as burn-in. Convergence was originally assessed using the potential scale reduction (PSR), taking PSR < 1.05 as a criterion, with a maximum of 75,000 iterations. Based on reviewers' comments, we reran the conditions for $N = 35$ with a fixed number of iterations, first 50,000 for each chain and then 100,000 for those replications that did not obtain a PSR < 1.05 (using the population values as starting values). We then assessed convergence by selecting those replications with a PSR < 1.1 [4] and manually checked a part of the traceplots. Given that the results did not differ substantially from the original results, we only took this approach for $N = 35$. [5]

Estimation error was assessed using (relative) bias and mean squared error (MSE). Bias was computed as $\frac{1}{S}\Sigma_{s=1}^{S}(\hat{\theta}_s - \theta)$ with S being the number of converged replications per cell, θ being the population value of that parameter, and $\hat{\theta}_s$ being the posterior median[6] for that parameter in a specific replication s. Relative bias was computed as $\frac{1}{S}\Sigma_{s=1}^{S}(\frac{\hat{\theta}_s - \theta}{\theta})$ and is only defined in those population conditions where $\theta \neq 0$. MSE was computed based on the true population value as: $\frac{1}{S}\Sigma_{s=1}^{S}(\hat{\theta}_s - \theta)^2$. We obtained the 95% coverage interval by computing how often the population value was contained in the 95% credible or confidence interval. In addition, to check how well the posteriors reproduced the sampling distributions, we investigated the

[4] Although less conservative, we used this cut-off because after 100,000 iterations some parameters had a PSR slightly above 1.05 while the traceplots indicated convergence based on eyeballing. An example is given in the supplemental material.

[5] Some priors and conditions were added later during the review process. Specifically, for $N = 150$ and $N = 500$, two vague proper and two noninformative improper priors were added later in those population conditions where the direct and indirect effect were manipulated, as well as the vague normal and EB2 prior in all population conditions. For these conditions, we simply ran all replications with 100,000 iterations and assessed convergence by selecting those replications with a PSR < 1.1, given that this strategy had proven correct for $N = 35$.

[6] We compared the posterior medians with the posterior means. Overall, these two posterior summaries did not differ substantially. However, in some replications the posterior mean for a specific parameter was extremely high. This can happen when the MCMC sampler samples some extreme values which have a large influence on the posterior mean, but not on the posterior median. Therefore, we used the posterior median as point estimate.

2.5% and 97.5% quantiles for every parameter. The 2.5% (97.5%) quantile was computed as the proportion of times that the lower (upper) bound of the 95% confidence/credible interval was higher (lower) than the population value. Ideally these should equal 2.5% and 97.5%, respectively. Finally, we looked at the Type 1 error rates for the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$. The ML results are included for comparison. All analyses were done in Mplus (version 7.2) and R, using the package MplusAutomation (Hallquist & Wiley, 2014).

Convergence

Table 2.4 shows the percentage of converged replications for each prior and sample size, averaged across the population values. For $N = 35$, the EB2 prior resulted in the highest convergence (98.9%), followed by the vague normal prior (94.1%). For all priors convergence generally increased with sample size and there was almost no convergence for the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$ when $N \leq 150$. This is not surprising since this prior is known to result in improper posteriors for variances of random effects in multilevel analysis (e.g., Gelman, 2006). Because of the severe nonconvergence under this improper prior we shall not consider it further in this paper. Note that this may imply that the vague proper priors $IG(\epsilon, \epsilon)$, with $\epsilon = 0.1, 0.01,$ or 0.001 can perform badly, as they approximate the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$. Specifically, traceplots of converged replications for these vague proper priors showed occassional peaks, resulting in relatively high posterior medians for those replications. We will only present the results in those population conditions with at least 50% convergence and we only consider the converged replications. Convergence percentages for each separate population condition are available in the supplemental material. The ML analysis always converged but often resulted in estimated negative variances. Specifically, in 53.9% of the replications at least one Heywood case occurred.

Table 2.4: Percentage converged replications for the default priors in the simulation study, averaged across population values.

N	Mplus default	$\pi(\sigma^2) \propto \sigma^{-1}$	$\pi(\sigma^2) \propto \sigma^{-2}$	IG(0.001, 0.001)	IG(0.01, 0.01)	IG(0.1, 0.1)	Vague normal	EB1	EB2
35	73.7	53.3	0	51.4	68.9	88.7	94.1	83.3	98.9
75	99.6	99.5	1.03	97.2	99.9	99.9	99.0	99.6	100
150	100	100	1.77	100	100	100	95.0	99.9	99.6
500	100	99.8	45.6	99.8	100	100	98.7	99.8	98.8

Note. N = sample size. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG(0.001, 0.001), IG(0.01, 0.01) , IG(0.1, 0.1) = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors.

Bias

Table 2.5 presents the relative bias, with the bias in brackets, for selected parameters in the model, for $N = 35$ and $N = 75$ averaged across population values. Results for all parameters in the model are available in the online supplemental material. Following (L. K. Muthén & Muthén, 2002), relative biases exceeding 10% are regarded as substantial and shown in bold. Given that the influence of the prior is greatest for small samples, we will focus on $N = 35$ and $N = 75$ in presenting the results and only mention the results for $N = 150$ and $N = 500$ briefly. The results for $N = 150$ and $N = 500$ can be found in the supplemental material.

For $N = 35$ the vague proper priors resulted in relative biases greater than 0.10 for most location parameters, followed by the EB priors. The noninformative priors (Mplus default, $\pi(\sigma^2) \propto \sigma^{-1}$) and the vague normal prior resulted in only a few parameters with substantial bias. ML performed best, with relative bias close to zero for all location parameters. For all priors, some parameters showed more bias than others. Specifically, the measurement intercepts ν_y and the structural intercepts α often resulted in large relative bias.

For $N = 75$, the bias decreased for all priors, resulting in only a few location parameters with relative bias exceeding 10% for the noninformative improper, vague proper, and vague normal priors. The EB priors performed worst, with 6 and 7 location parameters showing relative biases greater than 0.10, while ML again resulted in relative bias close to zero for all parameters. Again, the measurement intercepts ν_y showed most bias. For $N = 150$ only the vague proper and EB priors showed relative biases greater than 0.10 for some location parameters, while for $N = 500$ the relative bias was close to zero for all priors and location parameters.

For the variance parameters in the model, when $N = 35$, the vague proper prior IG$(0.1, 0.1)$, the vague normal prior, and the EB2 prior resulted in most cases with substantial bias, followed by the improper priors, and the vague proper prior IG$(0.01, 0.01)$. The vague proper prior IG$(0.001, 0.001)$ and the EB1 prior performed good, with only 3 variance parameters having relative biases greater than 0.10, and ML performed best with only 2 parameters with relevant bias. Generally, the estimated latent variable variances showed more bias then the estimated error variances, especially $\omega^2_{\zeta 65}$ which had relative bias greater than 0.10 for ML and all priors, except the IG$(0.001, 0.001)$ prior.

For $N = 75$, the vague normal and EB2 prior resulted in most biased variance parameters, followed by the Mplus default setting (i.e., the improper prior $\pi(\sigma^2) \propto 1$ combined with the N$(0, 10^{10})$ prior), then the vague proper priors IG$(0.001, 0.001)$ and IG$(0.1, 0.1)$, and next the EB1 prior. The improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ and the vague proper prior IG$(0.01, 0.01)$ performed well, with only 2 variances exceeding the relative bias of 10%, and ML performed best. Again, the estimates for the

latent variable variances were generally more biased than the estimates for the error variances. For $N = 150$, all methods resulted in some biased variance parameters, except the improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ and ML estimation. For $N = 500$, only IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, the Mplus default, and the EB1 prior resulted in some biased variance parameters.

Table 2.5: Relative bias and bias (in brackets) for each default prior and ML estimation, averaged across population values, for selected parameters

	γ_{60}	γ_{65}	b_{21}	$\gamma_{60} \cdot b_{21}$	α_{60}	α_{65}	λ_2^y	ν_2^y	$\omega^2_{\zeta 60}$	$\omega^2_{\zeta 65}$	ω^2_{D15}	σ^2_{y6}	σ^2_{x2}
						Sample size = 35							
Mplus default	0.044	0.026	-0.016	0.016	**0.13**	0.05	0.038	0.069	0.067	**1.207**	**0.332**	0.099	**0.427**
	(0.053)	(0.024)	(-0.013)	(0.017)	(-0.264)	(-0.118)	(0.047)	(-0.179)	(0.266)	(0.208)	(0.239)	(0.142)	(0.051)
$\pi(\sigma^2) \propto \sigma^{-1}$	0.051	-0.009	-0.003	0.035	**0.185**	0.006	0.026	0.078	0.014	**0.373**	0.011	**-0.227**	**0.134**
	(0.068)	(-0.002)	(-0.003)	(0.045)	(-0.336)	(-0.034)	(0.053)	(-0.181)	(0.159)	(0.082)	(0.133)	(-0.095)	(0.017)
IG$(0.001, 0.001)$	0.037	0	**0.271**	0.035	**0.137**	**0.611**	**0.468**	**1.411**	-0.019	0.005	-0.053	**-0.117**	**-0.111**
	(0.056)	(0.005)	(0.227)	(0.043)	(-0.278)	(-1.426)	(0.588)	(-3.683)	(-0.074)	(0.001)	(0.017)	(-0.17)	(-0.013)
IG$(0.01, 0.01)$	-0.05	-0.047	**0.191**	-0.042	**-0.185**	**0.364**	**0.752**	**2.015**	**-0.134**	**0.161**	**0.378**	**-0.208**	0.068
	(-0.048)	(-0.021)	(0.274)	(-0.027)	(0.251)	(-0.388)	(1.133)	(-2.444)	(-0.631)	(0.033)	(0.238)	(-0.097)	(0.006)
IG$(0.1, 0.1)$	-0.04	0.01	0.034	-0.042	**-0.15**	**0.119**	**0.155**	**0.417**	**-0.157**	**0.667**	**0.269**	**-0.238**	**0.339**
	(-0.046)	(0.004)	(0.046)	(-0.038)	(0.244)	(-0.163)	(0.255)	(-0.715)	(-0.728)	(0.128)	(0.25)	(-0.111)	(0.04)
Vague normal	-0.055	0.002	0.007	-0.069	**-0.162**	0.012	0.068	**0.101**	-0.042	**1.186**	**0.424**	0.099	**0.436**
	(-0.062)	(0.001)	(0.006)	(-0.067)	(0.328)	(-0.028)	(0.086)	(-0.264)	(-0.167)	(0.204)	(0.294)	(0.143)	(0.052)
EB1	-0.077	**-0.149**	0.011	-0.079	**-0.23**	**-0.2**	-0.019	**-0.162**	-0.04	**0.212**	-0.037	-0.089	0.036
	(-0.095)	(-0.103)	(0.009)	(-0.08)	(0.468)	(0.467)	(-0.023)	(0.422)	(-0.159)	(0.036)	(-0.002)	(-0.128)	(0.004)
EB2	-0.079	**-0.155**	-0.011	-0.099	**-0.236**	**-0.251**	-0.06	**-0.229**	0.084	**1.491**	**0.268**	0.096	**0.499**
	(-0.100)	(-0.109)	(-0.009)	(-0.102)	(0.479)	(0.585)	(-0.075)	(0.599)	(0.334)	(0.256)	(0.200)	(0.139)	(0.060)
ML	0.021	0.002	0.017	0.032	0.074	0.036	0.018	0.038	-0.061	**-0.246**	-0.078	**-0.107**	-0.033
	(0.030)	(0.006)	(0.015)	(0.035)	(-0.142)	(-0.088)	(0.024)	(-0.101)	(-0.235)	(-0.043)	(-0.053)	(-0.159)	(-0.004)
						Sample size = 75							
Mplus default	-0.016	0.036	-0.013	-0.036	-0.041	0.039	0.026	0.035	-0.003	**0.688**	**0.189**	0.016	**0.211**
	(-0.018)	(0.029)	(-0.01)	(-0.035)	(0.084)	(-0.092)	(0.033)	(-0.09)	(-0.01)	(0.118)	(0.142)	(0.023)	(0.025)
$\pi(\sigma^2) \propto \sigma^{-1}$	-0.018	0.038	-0.017	-0.041	-0.058	0.01	0.019	0.043	-0.016	**0.232**	**0.118**	-0.094	0.034
	(-0.034)	(0.017)	(-0.003)	(-0.04)	(0.166)	(-0.045)	(0.041)	(-0.123)	(-0.107)	(0.037)	(0.075)	(-0.002)	(0.001)
IG$(0.001, 0.001)$	-0.034	-0.002	0.007	-0.038	-0.094	0.004	0.06	**0.104**	-0.055	**-0.117**	0.027	-0.062	**-0.148**
	(-0.04)	(0.002)	(0.006)	(-0.037)	(0.192)	(-0.009)	(0.075)	(-0.273)	(-0.216)	(-0.02)	(0.03)	(-0.089)	(-0.018)
IG$(0.01, 0.01)$	-0.022	0	0.004	-0.022	-0.078	0.004	0.032	0.089	-0.052	-0.022	0.078	-0.097	-0.066
	(-0.039)	(0.008)	(0.007)	(-0.03)	(0.196)	(-0.053)	(0.056)	(-0.194)	(-0.243)	(0.002)	(0.036)	(-0.085)	(-0.005)
IG$(0.1, 0.1)$	-0.005	0.023	-0.006	-0.016	-0.018	0.01	0.026	0.075	-0.051	**0.334**	0.063	**-0.153**	**0.144**
	(-0.027)	(0.019)	(0)	(-0.031)	(0.143)	(-0.075)	(0.058)	(-0.194)	(-0.253)	(0.079)	(0.088)	(-0.091)	(0.018)
Vague normal	-0.036	0.025	-0.009	-0.051	**-0.113**	0.028	0.022	0.036	-0.004	**0.632**	**0.18**	0.032	**0.169**
	(-0.044)	(0.019)	(-0.008)	(-0.052)	(0.23)	(-0.065)	(0.028)	(-0.093)	(-0.017)	(0.109)	(0.135)	(0.046)	(0.02)
EB1	-0.068	**-0.102**	0.012	-0.061	**-0.205**	**-0.141**	-0.012	**-0.104**	-0.023	**-0.146**	-0.046	-0.069	-0.092
	(-0.087)	(-0.076)	(0.01)	(-0.067)	(0.416)	(0.33)	(-0.016)	(0.272)	(-0.09)	(-0.025)	(-0.015)	(-0.099)	(-0.011)
EB2	-0.068	-0.074	-0.01	-0.083	**-0.211**	**-0.131**	-0.027	**-0.128**	0.044	**0.761**	**0.133**	0.009	**0.259**
	(-0.088)	(-0.051)	(-0.008)	(-0.088)	(0.429)	(0.307)	(-0.034)	(0.333)	(0.174)	(0.131)	(0.108)	(0.013)	(0.031)
ML	-0.009	0.020	0.008	-0.003	-0.029	0.045	0.010	0.024	-0.029	**-0.135**	-0.041	-0.061	-0.005
	(-0.011)	(0.018)	(0.007)	(-0.003)	(0.000)	(-0.021)	(0.005)	(-0.017)	(-0.087)	(-0.043)	(0.000)	(-0.010)	(-0.038)

Note. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, IG$(0.1, 0.1)$ = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Values for the relative bias greater than 0.10 are shown in bold.

Overall, ML estimation performed best in terms of relative bias for both the variance and location parameters. This can be explained by the fact that ML estimation does not force separate variance components to be positive. For the location parameters, the vague proper priors and the EB priors performed worst and for

the variance parameters, the vague normal and EB2 prior performed worst. Of the Bayesian methods, the Mplus default setting performed best for the location parameters and the EB1 prior performed best for the variance parameters.

Mean squared error

Figure 2.3 shows for each prior and type of parameter the mean squared errors (MSEs) relative to the MSE of ML estimation per population value and parameter on the logarithmic scale, $\ln(\text{MSE}_{\text{Bayes}}/\text{MSE}_{\text{ML}})$. The results are categorized by structural regression coefficients, intercepts and latent mean, factor loadings, and variance parameters. Note that the vertical axis is truncated at $\ln(\text{MSE}_{\text{Bayes}}/\text{MSE}_{\text{ML}}) = 4$, excluding the extreme situations in which the MSE of the prior is more than $\exp(4) \approx 55$ times higher than the MSE of the maximum-likelihood estimate, which occurred for the vague proper priors. Tables with the numerical values for the MSE are available in the supplemental material. For $N = 35$, the vague proper priors performed worst, especially for the intercepts and factor loadings. This can be explained by the occassional extreme values drawn for these priors, as noted previously, and this unstability in the MCMC sampler is due to the fact that the vague proper priors approximate the problematic improper prior $\pi(\sigma^2) \propto \sigma^{-2}$. All other Bayesian methods resulted in smaller or approximately equal MSEs in comparison to ML estimation. In particular, the EB priors and the vague normal prior performed best for the structural regression coefficients. For $N = 75$, $N = 150$, and $N = 500$, there were hardly any clear differences between the MSEs using the different methods.

We can thus conclude that all methods perform similarly in terms of MSEs, except for the vague proper priors which performed considerably worse.

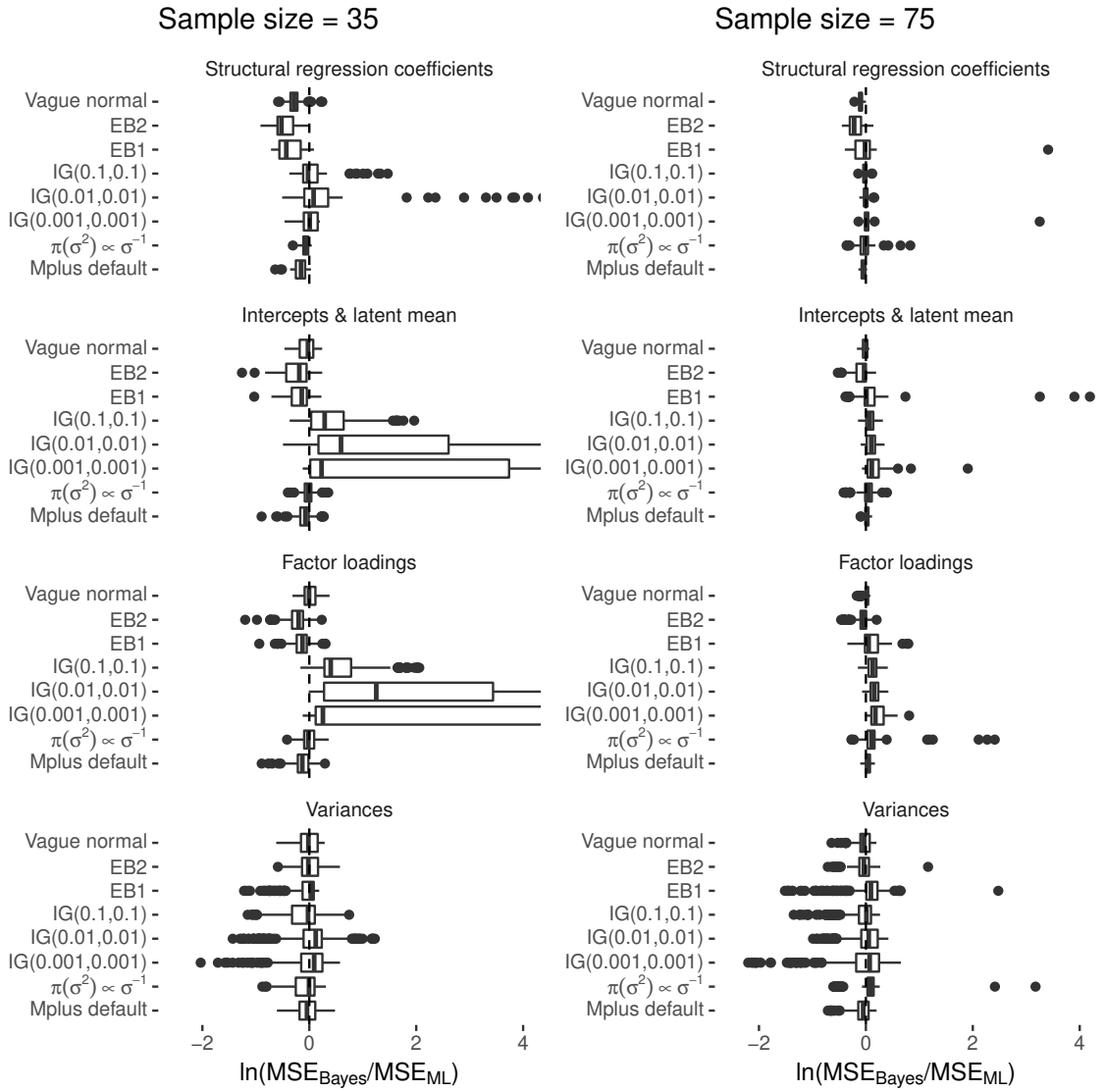

Figure 2.3: Mean squared error (MSE) for Bayesian estimation using different default priors divided by the MSE for maximum likelihood (ML) estimation on the natural logarithmic scale. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, IG$(0.1, 0.1)$ = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Vertical dashed lines indicate where the MSE for the Bayesian estimates equals the MSE for the ML estimates.

Coverage rates

Table 2.6 shows the coverage rates of the 95% confidence intervals and 95% Bayesian credible intervals for selected parameters. Coverage rates higher than 97.5% or lower than 92.5% are considered as substantially deviating from the desired

95%, and are marked in bold. Coverage rates for all parameters in the model are available in the online supplemental material. For $N = 35$, the EB1 and EB2 priors performed worst with low coverage for 13 and 12 location parameters, respectively, followed by the vague proper prior IG$(0.01, 0.01)$ which showed low coverage for 9 parameters. The vague normal prior showed coverage rates for 8 parameters that were too high, as did the Mplus default setting for 5 parameters. ML estimation resulted in coverage rates that were too low for 4 parameters. The improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ and the vague proper priors IG$(0.001, 0.001)$ and IG$(0.1, 0.1)$ performed best.

For $N = 75$, the vague proper prior IG$(0.001, 0.001)$ performed worst with low coverage rates for 16 location parameters, followed by the EB1 prior with low coverage rates for 14 parameters. The EB2 prior also showed low coverage for 8 parameters. The vague normal prior showed coverage rates for 2 parameters that were slightly too high. The other priors and ML estimation resulted in coverage rates between 92.5% and 97.5% for all parameters. For $N = 150$ and $N = 500$, all methods resulted in coverage rates for the location parameters close to 95%, except for the EB priors when $N = 150$.

We will now discuss the coverage rates for the variance parameters. [7] For $N = 35$, ML estimation performed worst with low coverage rates for 19 parameters. The EB1 prior also showed low coverage for 14 parameters, as did the EB2 prior for 9 parameters, while the Mplus default setting showed coverage rates for 11 parameters that were too high, as did the vague normal prior for 9 parameters. The improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ and the vague proper priors showed coverage rates for 6 or 8 parameters outside the range of 92.5% and 97.5%. For $N = 75$, coverage rates improved for most priors and ML estimation, except for the vague proper prior IG$(0.001, 0.001)$ and the EB1 prior which resulted in extreme values for 16 and 15 variance parameters, respectively. For $N = 150$, coverage rates for all variance parameters were close to 95% for the vague normal and EB2 prior. The EB1 prior performed worst, followed by the IG$(0.001, 0.001)$ and $\pi(\sigma^2) \propto \sigma^{-1}$ priors. For $N = 500$, ML estimation performed best and the EB1 prior performed worst, followed by the $\pi(\sigma^2) \propto \sigma^{-1}$ and IG$(0.001, 0.001)$ priors.

In summary, for location parameters the EB priors performed worst in terms of coverage rates, while for the variance parameters ML estimation, the vague proper

[7] We deleted some values for the coverage rates for the vague normal prior, the vague proper priors, the Mplus default, and the EB2 prior which were equal to zero. These values occurred only for Ω_D when the population value was zero. This happened because the lower bound of the credible interval was always greater than zero for these priors and thus a population value of zero is by definition never contained in the credible interval. For the EB1 prior and the improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ the lower bound did equal zero in several replications thereby resulting in a coverage not equal to zero in this situation. Note that, since the lower bound did not equal zero in all replications, the resulting coverage for these priors was low when the population values were equal to zero. For ML, the lower bound of the confidence interval can be negative.

prior IG(0.001, 0.001), and the EB1 prior performed worst. Across all parameters, the improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ and the vague proper prior IG(0.1, 0.1) performed best.

Table 2.6: Coverage rates of 95% Bayesian credible intervals and 95% confidence intervals for each default prior and ML estimation, averaged across population values, for selected parameters

	γ_{60}	γ_{65}	b_{21}	$\gamma_{60} \cdot b_{21}$	α_{60}	α_{65}	λ_2^y	ν_2^y	$\omega^2_{\zeta 60}$	$\omega^2_{\zeta 65}$	ω^2_{D26}	σ^2_{y4}	σ^2_{x2}
						Sample size = 35							
Mplus default	96.5	97.2	**98.2**	97	96.9	97.3	**98.2**	**97.7**	97.1	96.6	**98**	**98**	**97.7**
$\pi(\sigma^2) \propto \sigma^{-1}$	96.2	96.3	96.7	96.4	96.2	96	97.2	97	96.4	**98.2**	**83.3**	95.3	**98.8**
IG(0.001, 0.001)	94.8	95	95.4	96	94.9	95	95.4	95.8	94.9	**98.2**	**91.2**	**92.5**	**98.1**
IG(0.01, 0.01)	**91.8**	94.5	94	93.6	**91.9**	94.3	**90.8**	**91.3**	**86.2**	**98.8**	94	94	**99**
IG(0.1, 0.1)	93.4	96.3	96	94.9	93.6	95.8	94.5	94.7	**88.2**	**98.1**	95.7	95.3	**98.9**
Vague normal	96.7	**97.8**	**98.1**	97	97	**97.5**	**98.1**	**97.7**	95.1	96.9	**97.9**	**98.1**	**97.6**
EB1	94.9	**77.6**	**91.6**	93.4	95	**72.4**	**88.5**	**87.1**	**91.7**	**64.6**	**79.3**	**88.8**	**84.7**
EB2	96.3	**81.4**	96.0	95.0	96.9	**74.2**	**91.1**	**89.2**	95.0	95.3	**77.8**	**98.0**	96.4
ML	93	**90.2**	93.7	93.8	92.7	**90.9**	92.9	93.1	**87.4**	**90.7**	**90.4**	**92**	94
						Sample size = 75							
Mplus default	94.3	95.8	95.6	94.1	94.5	94.8	95.3	95.4	94.7	94.5	95.8	94.5	95.6
$\pi(\sigma^2) \propto \sigma^{-1}$	94.4	95.2	95.8	94.4	94.7	94.9	95	95.4	94.3	**97.6**	92.8	93.8	96.3
IG(0.001, 0.001)	**91.6**	**92.2**	92.7	**91.6**	**91.8**	**91.8**	**91.5**	**91.8**	**90.5**	94.7	**72.5**	**87.9**	**89.9**
IG(0.01, 0.01)	94.4	95	95.4	94.2	94.5	94.5	94.4	94.7	93.3	**97.7**	93	**91.5**	96.5
IG(0.1, 0.1)	94.4	95.8	95.3	94.3	94.4	95.2	94.4	94.8	93.5	97.1	95.1	**92**	**97.9**
Vague normal	95.3	96.6	97	95.1	95.6	96.3	96.2	95.9	95.2	96.3	96.4	94.9	96.3
EB1	92.8	**82.5**	**89.1**	**90.8**	93.3	**80.6**	**91.3**	**90.3**	92.6	**73.2**	**81.5**	**89.1**	**86.3**
EB2	94.2	**87.4**	95.1	92.5	94.8	**83.3**	92.9	**92.2**	94.6	94.1	**76.8**	95.1	95.2
ML	93.9	93.6	95.3	93.8	94.1	93.3	94.5	94.7	**92.0**	92.7	93.0	92.8	95.2

Note. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG(0.001, 0.001), IG(0.01, 0.01) , IG(0.1, 0.1) = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Coverage rates lower than 92.5% and higher than 97.5% are shown in bold.

Quantiles

Even in the case of perfect coverage rates of 95%, it may be that the underestimation of the interval estimate occurs much more often than overestimation (or the other way around). To assess this, we investigated the lower 2.5% and upper 97.5% quantiles. The quantiles were obtained by computing how often the lower 2.5% and upper 97.5% bounds of the credible/confidence intervals were above or below the true population value. Figure 2.4 shows the quantiles for $N = 35$ with the dashed lines indicating 2.5% and 97.5%. The results are categorized by structural regression coefficients, intercepts and latent mean, factor loadings, and variance pa-

rameters. The numerical values on which Figure 2.4 is based are available in the supplemental material. For the lower quantile, all priors resulted in quantiles close to the desired 2.5% except for the two EB prior settings in the case of intercepts and the vague proper priors in the case of factor loadings. For the upper quantile, the noninformative improper priors and the vague normal prior generally performed best with upper quantiles close to or slightly higher than 97.5%. For the structural regression coefficients and factor loadings, the EB priors performed worst, followed by ML estimation. For the variance parameters, the EB1 prior performed badly, as did ML estimation. For the intercepts, all priors and ML estimation resulted in quantiles close to the desired 97.5%.

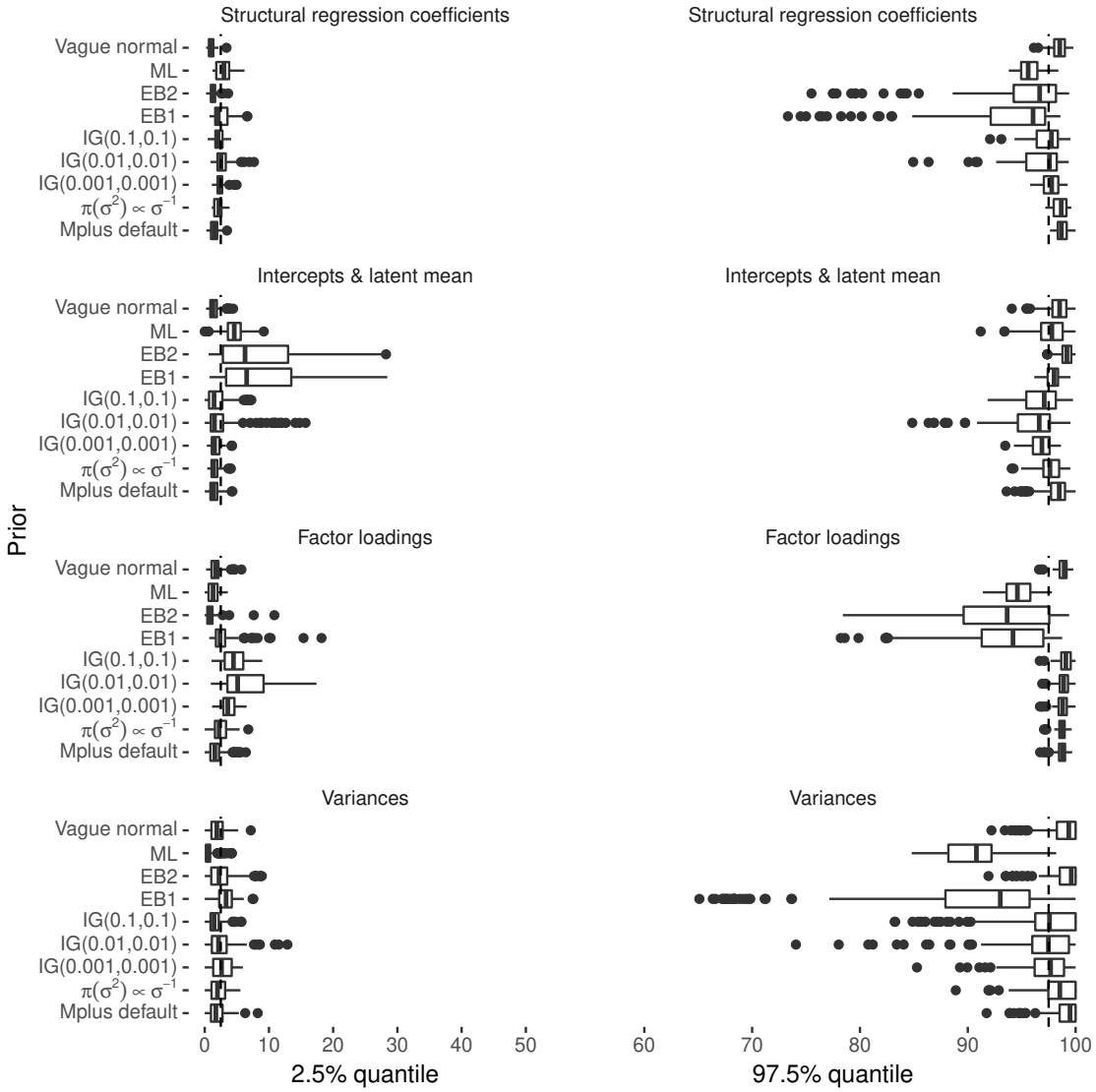

Figure 2.4: 2.5% and 97.5% quantiles for each default prior and maximum likelihood (ML) estimation for $N = 35$. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, IG$(0.1, 0.1)$ = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Vertical dashed lines indicate the desired 2.5% and 97.5%.

Figure 2.5 shows the quantiles for $N = 75$, which were all closer to the desired quantiles compared to $N = 35$. For the lower quantile, the EB priors resulted in too high quantiles for the intercepts. Note also the outliers for the vague proper prior IG$(0.001, 0.001)$ across parameters. For the upper quantile, the EB priors again

performed worst for the structural regression coefficients and loadings, followed by ML estimation. For the variance parameters, the EB1 prior and ML estimation also performed badly, while for the intercepts all priors and ML estimation resulted in quantiles close to the desired 97.5%. For $N = 150$, most priors resulted in quantiles close to the desired 2.5% and 97.5%, except for the vague normal and EB2 priors which were slightly higher or lower for some parameters. For $N = 500$, all methods generally resulted in correct quantiles.

To conclude, overall the EB priors performed worst in terms of quantiles, while the noninformative improper priors and the vague normal prior performed best.

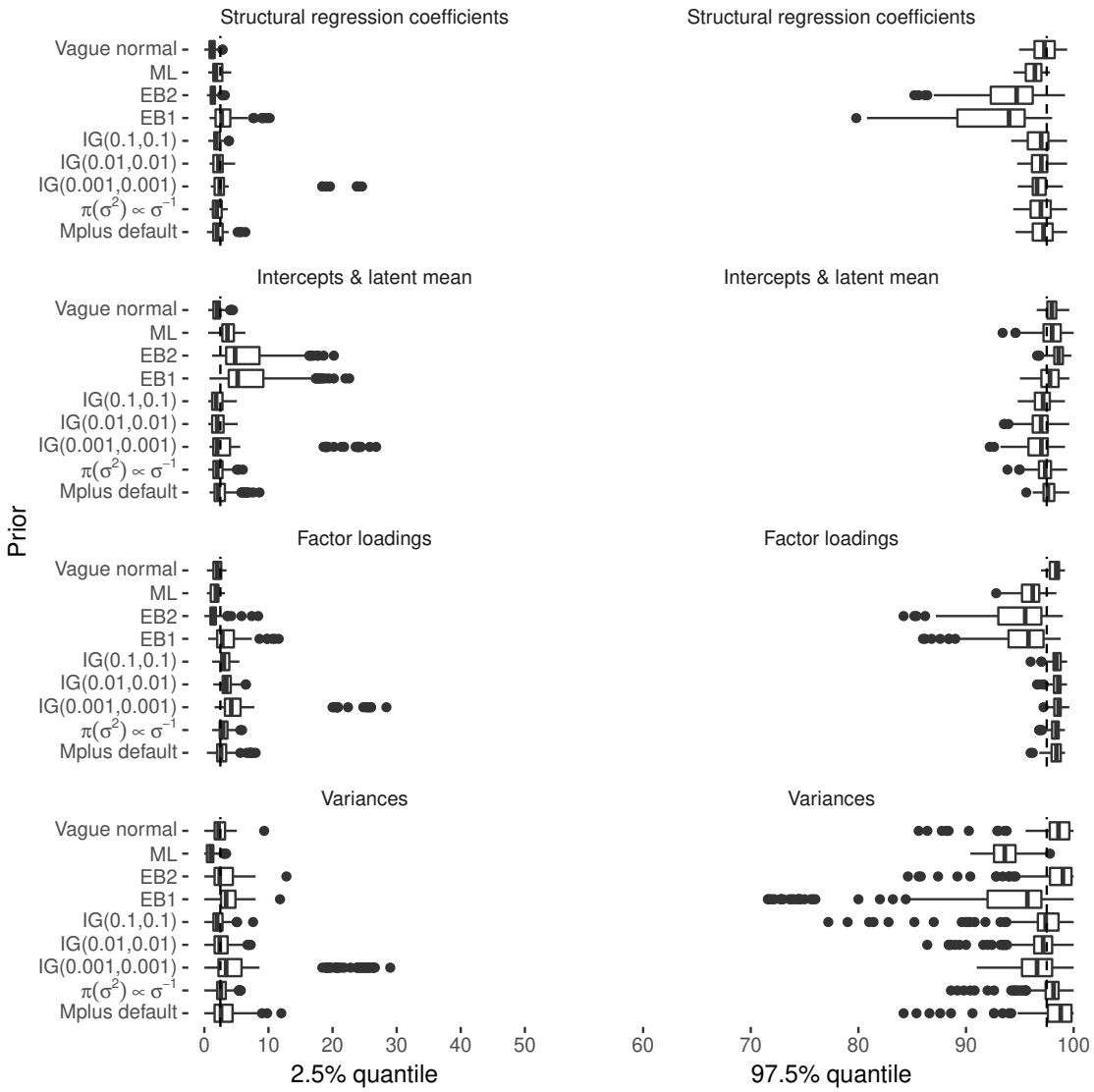

Figure 2.5: 2.5% and 97.5% quantiles for each default prior and maximum likelihood (ML) estimation for $N = 75$. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, IG$(0.1, 0.1)$ = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Vertical dashed lines indicate the desired 2.5% and 97.5%.

Direct and indirect effect

In practice, researchers often are mainly interested in those parameters related to the research question. In this model the parameters of substantive interest are

the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$. Figure 2.6 shows the MSEs of the direct and indirect effect for the different priors and ML estimation for $N = 35$ and $N = 75$. The figure shows that for $N = 35$, the vague normal and EB priors resulted in the smallest MSEs for the direct effect; the other methods showed slightly worse results. For the indirect effect and $N = 35$, the smallest MSEs were obtained using the EB2 prior, followed by the EB1 prior, and the vague normal prior. ML estimation performed worst.

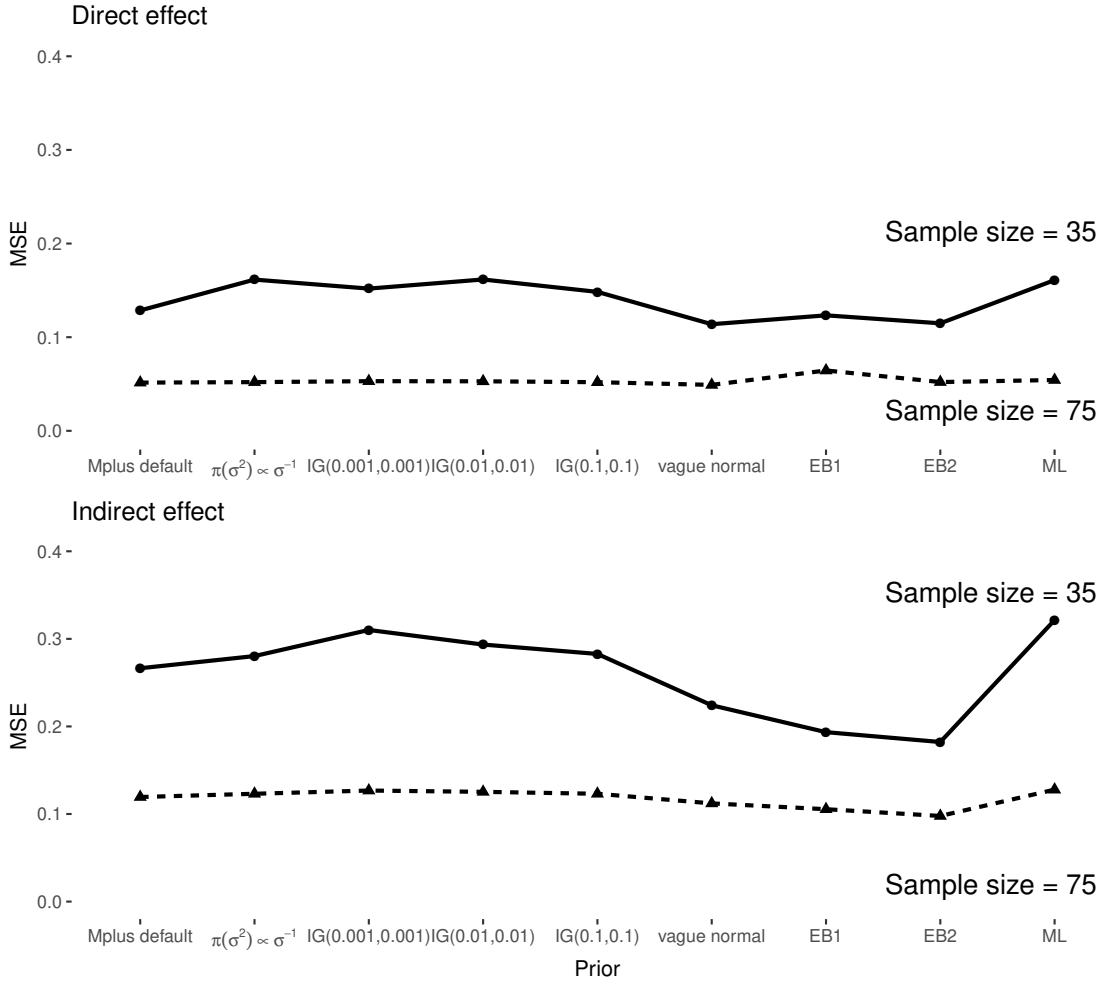

Figure 2.6: Mean squared error (MSE) for each default prior and maximum likelihood (ML) estimation for the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG$(0.001, 0.001)$, IG$(0.01, 0.01)$, IG$(0.1, 0.1)$ = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors.

Furthermore, we consider the type 1 error rates for the direct and indirect effect, i.e., the percentage of replications for which zero is not included in the 95% credible interval, or in the 95% confidence interval for the ML estimates, when the population value is zero. Table 2.7 shows the type 1 error rates for the different priors and different population conditions. For $N = 35$, the error rates for the direct effect were closest to the nominal 5% for the vague proper priors. For the Mplus default setting and the vague normal prior, the rates were too low. For the EB priors and for ML estimation the type 1 error rates were too high. For the indirect effect, the differences between priors were smaller and in general, most priors resulted in error rates close to 5%, except for the vague normal and EB2 priors which resulted in error rates that were too low. For $N = 75$, the EB priors again resulted in error rates that were too high for the direct effect, while the rates for ML estimation were closer to 5%. For the indirect effect, all priors and ML estimation generally resulted in error rates slightly higher than 5%. For $N = 150$ and $N = 500$ all type 1 error rates were generally close to 5%, except for the direct effect for the vague proper priors and EB priors when $N = 150$.

Table 2.7: Type 1 error rates for the direct and indirect effect, for the different default priors in different population conditions

Parameter	Mplus default	$\pi(\sigma^2) \propto \sigma^{-1}$	IG(0.001, 0.001)	IG(0.01, 0.01)	IG(0.1, 0.1)	Vague normal	EB1	EB2	ML
				Sample size = 35					
Direct effect	2.6	NA	5.9	4.1	3.5	2.0	15.8	12.1	10.9
Indirect effect	4.2	4.0	NA	5.1	5.2	2.9	4.4	2.8	5.1
				Sample size = 75					
Direct effect	3.5	4.0	4.6	4.3	3.4	3.0	13.7	9.7	6.0
Indirect effect	6.1	6.1	6.5	6.4	6.1	5.4	7.1	6.1	6.3

Note. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG(0.001, 0.001), IG(0.01, 0.01) , IG(0.1, 0.1) = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal and EB priors. Some results are not available (NA) for conditions that did not have at least 50% convergence.

Based on these results, we can conclude that there is not one default prior that performed consistently better than the other priors or than ML estimation across all parameters or outcomes, especially in small samples. When looking across all parameters for $N = 35$, ML estimation performed best in terms of bias; in terms of MSE, the EB priors performed best. However, both the EB priors and ML estimation performed badly in terms of coverage and quantiles. The vague proper priors performed worst in case of bias and MSE but best in terms of coverage and type 1 error rates for the direct effect and indirect effect. Overall, the noninformative improper priors $\pi(\sigma^2) \propto 1$ and $\pi(\sigma^2) \propto \sigma^{-1}$ performed best across all parameters and

outcomes with good coverage rates and quantiles, and average bias and MSE. One disadvantage of the improper priors in small samples is their high nonconvergence percentage, especially for $\pi(\sigma^2) \propto \sigma^{-1}$.

Of the noninformative improper priors, the Mplus default setting (i.e., $\pi(\sigma^2) \propto 1$ for variances coupled with $N(0, 10^{10})$ for location parameters) performed best in terms of bias for the location parameters, but worse than $\pi(\sigma^2) \propto \sigma^{-1}$ for the variance parameters. In terms of coverage rates, the Mplus default setting outperformed the $\pi(\sigma^2) \propto \sigma^{-1}$ prior, especially for the variance parameters. For the MSE and quantiles, they did not differ substantially. When considering the direct and indirect effect, the parameters of practical interest, both noninformative improper priors performed good in terms of bias and coverage. Finally, although the type 1 error rate for the direct effect was slightly too low for the Mplus default setting when $N = 35$, this result was not available for the $\pi(\sigma^2) \propto \sigma^{-1}$ prior, due to high nonconvergence. Based on these differences in performance between the noninformative improper priors, we thus recommend the Mplus default priors as general choice for BSEM, with the important observation that this setting does not perform perfectly.

Even though the different default priors that were investigated in this section are routinely used in practice, their performance varies greatly across conditions. Therefore a prior sensitivity analysis is highly recommendable, especially in small samples when clear prior information is absent. The next section will provide guidelines on how to perform such an analysis in default BSEM.

2.5 A practical guide to prior sensitivity analysis

Given the results of the simulation study and in line with recommendations by (B. O. Muthén & Asparouhov, 2012, p. 320), prior sensitivity analysis is an important step in BSEM. The goal of a prior sensitivity analysis is to assess whether the results of the BSEM analysis are influenced by the specific default prior that is used. When the conclusions are similar using the different default priors, we can be confident that the results are reliable and robust to default prior specification. On the other hand, if the different default approaches result in substantially different conclusions, some care must be taken regarding the reliability of the results.

A prior sensitivity analysis can be conducted by rerunning the analysis with different choices for the prior. Due to the large number of parameters, possible prior choices, and possible settings for each choice, conducting a sensitivity analysis can become quite involved. Nevertheless, prior sensitivity analysis is particularly relevant in the context of SEM. Due to the complex relationships inherent in structural equation models, a prior on a specific parameter can indirectly influence other parts of the model as well. Consequently, the effects of the prior for different parameters

may cancel out, but can also accumulate. Moreover, some parameters are more influenced by their prior distribution (e.g., variances of latent variables) compared to other parameters (e.g., residual variances). If a parameter is highly influenced by its prior distribution, this influence can carry through the model and affect other parameters as well. Depaoli and van de Schoot (2017) provided guidelines on conducting prior sensitivity analyses for Bayesian analyses with informative priors. The goal of this section is to provide a step-by-step guide on how to conduct a prior sensitivity analysis in BSEM using default priors. This analysis is recommended when prior information is weak, or when a researcher prefers to exclude external information in the statistical analysis. We will illustrate the guidelines on the democracy and industrialization model from Section 2.

Step 1: Decide which parameters to investigate

The first step in conducting a sensitivity analysis for structural equation models is to decide which parameters to focus on. Although it is important to change the prior on each parameter, there are generally only a few parameters of substantive interest (e.g., the direct and indirect effect in the model considered throughout this paper). Therefore, we recommend to focus primarily on the parameters of substantive interest in determining the sensitivity to the prior. Which parameters are of interest will, of course, vary across different applications. In addition to the parameters of substantive interest we recommend to focus on the latent variable variances in the model as well since these are generally most sensitive to the choice of the prior. These variance parameters can therefore unduly influence other parameters as well. In this first step, it is also helpful to consider which magnitudes of changes in the parameter values would constitute meaningful differences in the parameters. These magnitudes will be used in Step 4 to determine when a parameter is sensitive to the choice of the prior.

Step 2: Decide which priors to include

The second step consists of deciding which priors to include. Software such as Mplus limits the choice of possible priors by allowing a limited set of prior choices, such as normal priors for location parameters (e.g., intercepts, regression coefficients) and inverse Gamma priors for variance parameters. Given the large number of parameters in the model, it is infeasible to alter the prior for each parameter one at a time. In addition, it is more realistic to change the prior for all parameters in the model simultaneously because in general researchers will specify the same type of default prior for all parameters in the model.

Based on the results of the simulation study, we recommend to include the following default priors in the prior sensitivity analysis: the noninformative improper priors $\pi(\sigma^2) \propto \sigma^{-1}$ and $\pi(\sigma^2) \propto 1$, and the vague proper priors $IG(\epsilon, \epsilon)$ with $\epsilon = 0.001, 0.01,$ and 0.1 for variance parameters, combined with the vague proper prior $N(0, 10^{10})$ for location parameters; the vague normal prior; and the EB priors. Note that it is important to consider multiple values for ϵ when considering the vague proper prior $IG(\epsilon, \epsilon)$, since the choice of ϵ can have a large influence on the results. When the results are not robust to the exact choice of ϵ, we do not recommend using these priors for drawing substantive conclusions. On the other hand, when the results are robust to the choice of ϵ, the results are reliable for drawing substantive conclusions. We do not recommend to include the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$, due to its severe nonconvergence.

Furthermore, when prior knowledge is available, a researcher can use an informative prior. This prior can be specified by choosing the hyperparameters in such a way that the resulting prior has high probability on those parameter values deemed plausible by previous research or by an expert in the field. The challenge in specifying informative priors is to specify the hyperparameters such that the prior probability that the parameter falls in a plausible parameter region equals a certain percentage, e.g., 95%. When informative priors are used, we follow the recommendation of Depaoli and van de Schoot (2017) to compare the results of the informative priors to results obtained using default priors.

Step 3: Technical implementation (Mplus)

The R package MplusAutomation (Hallquist & Wiley, 2014) can be used to automatically create and run the Mplus input files for the analyses with different priors. Subsequently, the results of all analyses can be read into R simultaneously. In the supplemental material we provide the code for our sensitivity analysis, which can be used as a template for a prior sensitivity analysis using MplusAutomation.

One issue when conducting the analyses in an automatic way is how to assess convergence. When using Markov Chain Monte Carlo (MCMC) sampling, it is important to ensure that the chains converge to the posterior distribution. Mplus provides an automatic criterion based on the potential scale reduction (PSR). Sampling will continue until the cut-off for the PSR defined in the BCONVERGENCE option is reached or before that if the maximum number of iterations is reached (specified through the BITERATIONS option). The maximum number of iterations should depend on the model under consideration, with more complex models requiring a larger number of iterations and preliminary analyses can be conducted to get an indication of the required number of iterations. In addition, when using the PSR it is recommended to rerun the analysis with twice as many iterations to

avoid preliminary fulfillment of the PSR criterion. More information on the PSR can be found in Gelman and Rubin (1992) or the Mplus User guide (L. K. Muthén & Muthén, 1998-2012). A second option is to not rely on the automatic criterion to determine the number of iterations, but to specify a fixed number of iterations (using the FBITERATIONS options) and subsequently assessing convergence diagnostics, such as the PSR. Again, preliminary analyses can be conducted to get an indication of the required number of iterations. Note that there exist other methods to automatically assess convergence, for example, blavaan has a setting which relies on the PSR in combination with Raftery and Lewis (1991) convergence diagnostic to determine the number of draws. Regardless of which automated criterion is used to assess convergence, it is highly recommended to check the traceplots of the posterior draws for all parameters.

Step 4: Interpretation of the results

The marginal posterior distributions for each parameter in the model can be summarized in different ways. As Bayesian point estimates the mean, median, or mode can be used. By default, Mplus provides the posterior median, which is also the summary we used. In addition, we considered the 95% credible interval. As noted in Step 1, in order to conclude whether the results are sensitive to the prior, the researcher must first decide what constitutes a meaningful difference in parameters of interest, based on the application at hand. In other words, boundaries must be specified for the changes in results across the priors; if a change in a parameter exceeds this boundary, the parameter can be classified as sensitive. Changes can be evaluated by comparing the results obtained with a specific prior to the results obtained with the original prior distribution. To define a meaningful boundary, it may be helpful to set the bounds on the standardized estimates, which are generally easier to interpret. In addition, because the standardized estimates automatically include the scale of the variables, only the sensitivity of the latent mean, intercept, loading, and regression parameters needs to be considered. However, other options are possible, for example the threshold can be based on qualitative differences. One such option is to classify a parameter as sensitive if a different prior results in a different sign of the estimate, or if the EPC (Saris, Satorra, & van der Veld, 2009), or EPC-interest (Oberski, 2014) exceeds a certain cut-off. The EPC or "expected parameter change" estimates the change in a parameter when relaxing a constraint, while the EPC-interest estimates the expected change in the parameter of interest. Sensitivity with respect to other outcomes may also be evaluated, such as whether the credible interval includes zero or not; or whether model fit measures, such as the posterior predictive p-value (Gelman, Meng, & Stern, 1996), exceed a threshold such as 0.05.

The results of the prior sensitivity analysis can fall into one of three categories: (1) the results are not sensitive to the choice of the prior; (2) the results obtained using default priors do not vary, but these results differ from the result obtained using informative priors; and (3) the results vary across all priors, both default and informative. In scenario (1) we can conclude that the results are robust to the choice of the prior. In scenario (2) the information embedded in the informative prior influences the results, as would be expected. As noted by Berger (2006), subjective prior elicitation is difficult and can generally only provide certain characteristics of the prior (e.g., the location), whereas other features (e.g., the parametric form) are typically chosen in a convenient way. Thus, in scenario (2), the researcher should be certain that the chosen informative prior is an accurate reflection of one's prior belief, in terms of the prior guess (i.e., the prior mean) and the prior uncertainty (i.e., the prior variance and prior's distributional form). If the certainty about the informative prior cannot be warranted, the analysis based on default priors is recommended for substantive conclusions. If the informative prior is an accurate representation of the prior beliefs, the results from this prior can be used for final conclusions, while the results of the default analyses can be used as a reference.

In scenario (3) there is an additional difficulty that the results are also not robust to the different default priors. This may occur when the sample is relatively small and the prior is (possibly unintentionally) relatively informative. In this situation, one option is to collect more data.The advantage of Bayesian analysis is that we can simply collect additional data and combine it with the original data, whereas classical methods, such as confidence intervals and p-values, require a fixed sampling plan before data analysis (e.g., Robert, 2007, p. 23). Thus, the researcher can continue to collect additional observations until the results are no longer sensitive to the priors. However, it is not always feasible or possible to collect more data. In that case, we recommend that the researcher reports all results or the range of results obtained using the different default priors. The range can be computed by first combining all posterior draws from the different priors, and then computing the median and the lower and upper bounds of the 95% credible interval of the combined set of posterior draws. In addition, differences between default priors can be examined graphically, for example using boxplots such as those shown in Figure 2.7. The interpretation should then focus on how the substantive conclusions (e.g., the effect size of the indirect effect) vary across the default priors. This mirrors the recommendation of Leamer (1983) and relates to robust Bayesian analysis (Berger, 2006), where the results from multiple prior distributions are combined to obtain a range of results.

Note that the three scenarios can occur for different parameters. If results between priors vary only for the nuisance parameters and not for the parameters of

interest (i.e., the direct and indirect effect), reliable conclusions can be drawn about these parameters of interest. Only when the researcher wishes to draw conclusions about the complete fitted model, sensitivity of the nuisance parameters to the priors should be taken into account.

2.6 Empirical application: democracy and industrialization data

We applied these steps to the original data from the democracy and industrialization application, which has a sample size of 75. In addition, we took the first 35 observations of the original data to illustrate a prior sensitivity analysis in a situation where the results are quite sensitive to the choice of the prior. For comparison, we also include the ML estimates in the results.

Step 1

In this specific application, the parameters of substantive interest are the direct effect γ_{65} and the indirect effect $\gamma_{60} \cdot b_{21}$. We decided that a standardized change of 0.1 would constitute a meaningful difference in the parameters. Note that the variables have been standardized with respect to the variances of both $\boldsymbol{y}$ and $\boldsymbol{x}$.

Step 2

We include in our sensitivity analysis the noninformative improper prior $\pi(\sigma^2) \propto \sigma^{-1}$ (combined with the vague proper prior $\mathrm{N}(0, 10^{10})$ for location parameters), the Mplus default setting (i.e., $\pi(\sigma^2) \propto 1$ for variance parameters and $\mathrm{N}(0, 10^{10})$ for location parameters), the vague proper priors (with $\epsilon = 0.1, 0.01,$ and 0.001, combined with the vague proper prior $\mathrm{N}(0, 10^{10})$ for location parameters), the vague normal prior, and the EB priors. The Mplus default prior setting is used as baseline to which the other priors are compared. In general, however, the original prior distribution will serve as baseline. Although we do not recommend the use of the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$, we did include this prior in the sensitivity analysis for illustrative purposes.

In addition, informative priors are available in Dunson et al. (2005) for this model based on expert knowledge, which we included in our prior sensitivity analysis (see the online supplemental material). We assume that these priors reflect the available prior knowledge and the certainty about this knowledge correctly.

Step 3

We ran the analyses using MplusAutomation. All files for our analyses are available in the supplemental material. We did not rely on the PSR criterion to determine the amount of iterations, but instead specified a fixed number of 75,000 iterations (through the FBITERATIONS option). For each analysis, we checked whether the PSR < 1.1 for all parameters and we eyeballed each traceplot to ensure convergence. An example of a traceplot illustrating convergence and the corresponding estimated posterior densities is available in the online supplemental material.

Step 4

To assess sensitivity, we compared the standardized median for each prior with the standardized median obtained when using the Mplus default prior settings (the baseline). The Mplus default settings correspond to a normal prior, $\mathrm{N}(0, 10^{10})$, for means and regression parameters and an improper prior, $\pi(\sigma^2) \propto 1$, for variances, implemented as an inverse Gamma prior, $\mathrm{IG}(-1, 0)$. Note that, in practice, the original prior distributions will serve as baseline. As noted in Step 1, a standardized change of 0.1 would constitute a meaningful difference. Consequently, if the standardized median of a prior deviated more than 0.1 from the standardized median obtained under the baseline, we concluded that the results are sensitive to the prior. We will first discuss the results for the original data ($N = 75$) followed by the results for $N = 35$.

Results original data ($N = 75$)

Table 2.8 shows the standardized and unstandardized ML estimates and posterior medians for the direct effect for each analysis and sample size, as well as the 95% confidence and credible intervals. Standardized estimates that deviate more than 0.1 from the estimates obtained with the Mplus default setting are presented in bold. It is clear that for $N = 75$, none of the estimates exceed the cut-off and thus we can conclude that the direct effect is not sensitive to the prior. There are differences between the priors in terms of credible intervals. Specifically, for the informative prior, the lower bound is negative while it is positive for all other priors and ML estimation. Consequently, a test for the direct effect using the informative prior would result in the conclusion that the effect is not substantially different from zero, while the other priors and ML estimation would lead to this conclusion. Note that the informative prior results in the smallest credible interval because of the additional information added through the prior. In addition, the EB priors have slightly smaller credible intervals compared to the other default priors, because the EB priors include more prior information than the default priors.

Table 2.8: Standardized and unstandardized point estimates and 95% confidence and credible intervals for the direct effect γ_{65} in the prior sensitivity analysis

Prior	Standardized estimate	Unstandardized estimate	Lower bound 95% CI	Upper bound 95% CI	Width 95% CI
			Sample size = 35		
Mplus default	0.299	1.137	0.132	2.193	2.061
$\pi(\sigma^2) \propto \sigma^{-1}$	0.284	1.052	0.090	2.087	1.997
IG(0.001, 0.001)	0.270	0.990	0.074	2.041	1.967
IG(0.01, 0.01)	0.278	1.019	0.059	2.029	1.970
IG(0.1, 0.1)	0.283	1.052	0.088	2.053	1.965
Vague normal	0.293	1.085	0.114	2.086	1.972
EB1	0.270	0.975	0.160	1.741	1.581
EB2	0.274	0.997	0.137	1.812	1.675
Informative	**0.090**	0.225	-0.427	0.791	1.218
			Sample size = 75		
Mplus default	0.183	0.574	0.098	1.092	0.994
$\pi(\sigma^2) \propto \sigma^{-1}$	0.177	0.554	0.082	1.046	0.964
IG(0.001, 0.001)	0.174	0.552	0.089	1.023	0.934
IG(0.01, 0.01)	0.175	0.549	0.078	1.031	0.953
IG(0.1, 0.1)	0.177	0.555	0.084	1.052	0.968
Vague normal	0.182	0.569	0.096	1.080	0.984
EB1	0.153	0.475	0.055	0.873	0.818
EB2	0.158	0.488	0.064	0.910	0.846
Informative	0.109	0.288	-0.126	0.678	0.804
ML	0.182	0.572	0.114	1.030	0.916

Note. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG(0.001, 0.001), IG(0.01, 0.01) , IG(0.1, 0.1) = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal, EB, and informative priors. Standardized estimates deviating more than 0.1 from the estimate obtained under the Mplus default prior settings are shown in bold.

Table 2.9 shows the point estimates and 95% confidence and credible intervals for the indirect effect for each analysis and sample size. None of the estimates exceed the cut-off of 0.1 and thus the indirect effect is not sensitive to the choice of the prior. The confidence and credible intervals for the indirect effect show less variation compared to the direct effect so that conclusions based on interval testing for the indirect effect would not differ across the Bayesian methods or ML estimation. Only the width of the credible intervals differ, with the informative prior showing the smallest credible intervals, followed by the EB priors.

Table 2.9: Standardized and unstandardized point estimates and 95% confidence and credible intervals for the indirect effect $\gamma_{60} \cdot b_{21}$ in the prior sensitivity analysis

Prior	Standardized estimate	Unstandardized estimate	Lower bound 95% CI	Upper bound 95% CI	Width 95% CI
			Sample size = 35		
Mplus default	0.430	1.572	0.556	2.938	2.382
$\pi(\sigma^2) \propto \sigma^{-1}$	0.446	1.599	0.580	2.962	2.382
IG(0.001, 0.001)	0.459	1.627	0.603	2.992	2.389
IG(0.01, 0.01)	0.456	1.622	0.609	2.996	2.387
IG(0.1, 0.1)	0.452	1.622	0.623	2.976	2.353
Vague normal	0.422	1.505	0.537	2.798	2.261
EB1	0.405	1.417	0.537	2.546	2.009
EB2	0.387	1.353	0.473	2.487	2.014
Informative	0.461	1.125	0.593	1.897	1.304
			Sample size = 75		
Mplus default	0.385	1.191	0.522	1.982	1.460
$\pi(\sigma^2) \propto \sigma^{-1}$	0.394	1.221	0.548	2.017	1.469
IG(0.001, 0.001)	0.399	1.241	0.569	2.050	1.481
IG(0.01, 0.01)	0.397	1.229	0.561	2.035	1.474
IG(0.1, 0.1)	0.393	1.208	0.553	2.011	1.458
Vague normal	0.383	1.177	0.522	1.950	1.428
EB1	0.376	1.150	0.592	1.801	1.209
EB2	0.364	1.108	0.554	1.751	1.197
Informative	0.377	0.988	0.582	1.482	0.900
ML	0.396	1.242	0.542	1.941	1.399

Note. ML = maximum likelihood estimation. EB1 = Empirical Bayes prior location and variance parameters. EB2 = EB prior location parameters combined with $\pi(\sigma^2) \propto \sigma^{-1}$ prior variance parameters. Vague normal = $\pi(\sigma^2) \propto 1$ prior for variance parameters combined with the normal $N(0, 1000)$ prior for measurement intercepts and the normal $N(0, 100)$ prior for the other location parameters. Mplus default = $\pi(\sigma^2) \propto 1$ combined with the normal $N(0, 10^{10})$ prior. $\pi(\sigma^2) \propto \sigma^{-1}$ = noninformative improper priors variance parameters. IG(0.001, 0.001), IG(0.01, 0.01) , IG(0.1, 0.1) = vague proper priors variance parameters. Location parameters have the normal $N(0, 10^{10})$ prior, except for the vague normal, EB, and informative priors. Standardized estimates deviating more than 0.1 from the estimate obtained under the Mplus default prior settings are shown in bold.

Results subset original data ($N = 35$)

For the subset of 35 observations from the original data, the ML analysis led to empirical weak identification and inadmissible estimates due to the small sample size and thus these results are excluded. The standardized and unstandardized posterior medians, and 95% credible intervals for the direct effect are presented in Table 2.8. The standardized median for the informative prior is shown in bold, indicating that this estimate differs more than 0.1 from the estimate obtained under the Mplus default setting. In addition, the informative prior is the only prior resulting in a negative lower bound of the 95% credible interval. Thus, for $N = 35$, the direct effect is sensitive to the choice of the prior. The results for the indirect effect are presented in Table 2.9. Again, the indirect effect is less sensitive to the choice of

the prior. None of the standardized estimates exceed the cut-off and the only clear difference between the priors is that the width of the credible interval is smallest for the informative prior, followed by the EB priors.

Conclusions from the prior sensitivity analysis

Based on the scenarios described in Step 4 of the prior sensitivity guide, we can thus conclude that for $N = 75$, the estimates of the parameters of interest (i.e., the direct and indirect effect) are robust to the choice of the prior (scenario (1)). However, the credible interval for the direct effect included zero for the informative prior whereas it did not include zero for the default priors. Thus, when testing the direct effect, we find ourselves in scenario (2). The same holds for $N = 35$, where both the estimate and credible interval of the direct effect were sensitive to the informative prior. Therefore, careful consideration of the informative prior for the direct effect is necessary. The informative prior for the direct effect was the normal prior $N(0.5, 2)$, which results in 95% prior probability on the interval $(-3.43, 4.42)$ (see the online supplemental material). Compared to the default priors, which are more spread out, the informative prior shrinks the estimate for the direct effect towards the prior mean, resulting in a smaller estimate. If the informative prior has been specified with care and accurately reflects the prior beliefs (we assume this was the case), the results obtained with the informative prior can be used for substantive conclusions, which implies no significant direct effect. The default analysis, which suggest a significant direct effect, can be reported as a reference analysis to show that the information in the data implies a significant direct effect.

For both sample sizes, the nuisance parameters were sensitive to the default priors as well. Thus, if the goal of the analysis is to draw conclusions about the full model, scenario (3) is applicable. If no informative priors were specified, and if it is not possible to collect more data, the researcher should consider and report the (range of) results from all default priors. By combining the posterior draws from all default priors and computing the median and bounds of the 95% credible interval, we can obtain a range for all parameters, which is reported in Table 2.10. Some of the credible intervals based on all posterior draws are very wide. This common behavior of a robust Bayesian analysis (e.g., Berger, 2006) can be explained by the fact that there is very little information in the data to fit the relatively complex SEM model.

Additionally, we can examine the differences between the default priors graphically, for example by plotting the standardized posterior medians for each parameter, as is done in Figure 2.7 for the structural intercept α_{60}. From Figure 2.7 we can see that for $N = 35$, the estimated medians vary from -1.4 to -2 and the researcher should further examine these differences between the priors. For example, in this

case, the smallest estimates are obtained using the EB priors, whereas the improper and vague proper priors generally result in estimates close to -2, and the vague normal prior lies in between. Of the default priors, the EB priors are most informative, as they include information regarding the ML estimates. The improper and vague proper priors are least informative, since they have the largest posterior variance, and the vague normal prior lies in between. Thus, for more informative default priors, the estimate for α_{60} becomes smaller. If informative priors were specified and scenario (3) is applicable, the researcher should carefully consider each of the informative priors and if in doubt whether the prior accurately reflects the prior belief, the researcher should consider the results of all default priors.

Table 2.10: Posterior medians and lower and upper bounds of the 95% credible interval based on all posterior draws from sensitivity analyses with default priors.

Parameter	Lower bound 95% CI	Median	Upper bound 95% CI	Lower bound 95% CI	Median	Upper bound 95% CI
		Sample size = 35			Sample size = 75	
γ_{60}	0.211	0.538	0.771	0.211	0.428	0.617
γ_{65}	0.032	0.281	0.530	0.032	0.172	0.323
b_{21}	0.541	0.780	0.953	0.541	0.885	0.971
$\gamma_{60} \cdot b_{21}$	0.114	0.420	0.735	0.114	0.379	0.599
λ_1^y	0.609	0.845	0.955	0.609	0.845	0.919
λ_2^y	0.492	0.737	0.886	0.492	0.692	0.811
λ_3^y	0.456	0.721	0.871	0.456	0.717	0.826
λ_4^y	0.637	0.842	0.951	0.637	0.831	0.909
λ_5^y	0.683	0.850	0.936	0.683	0.802	0.882
λ_6^y	0.528	0.758	0.887	0.528	0.726	0.832
λ_7^y	0.668	0.839	0.931	0.668	0.818	0.894
λ_8^y	0.483	0.718	0.861	0.483	0.812	0.892
λ_1^x	0.887	0.950	0.986	0.887	0.917	0.957
λ_2^x	0.935	0.985	1.000	0.935	0.973	0.998
λ_3^x	0.742	0.868	0.935	0.742	0.867	0.919
μ_ξ	5.781	7.775	10.120	5.781	7.415	9.038
α_{60}	-3.750	-1.694	0.898	-3.750	-0.723	0.972
α_{65}	-3.458	-1.791	-0.179	-3.458	-0.999	0.022
ν_2^y	-1.959	-0.741	0.015	-1.959	-0.600	-0.086
ν_3^y	-0.637	0.258	1.091	-0.637	0.213	0.768
ν_4^y	-1.575	-0.308	0.395	-1.575	-0.692	-0.226
ν_6^y	-1.600	-0.850	-0.223	-1.600	-0.857	-0.370
ν_7^y	-0.649	0.038	0.663	-0.649	-0.108	0.384
ν_8^y	-1.076	-0.328	0.359	-1.076	-0.704	-0.238
ν_2^x	-5.936	-4.254	-3.001	-5.936	-4.060	-3.095
ν_3^x	-5.515	-3.969	-2.627	-5.515	-3.901	-2.955

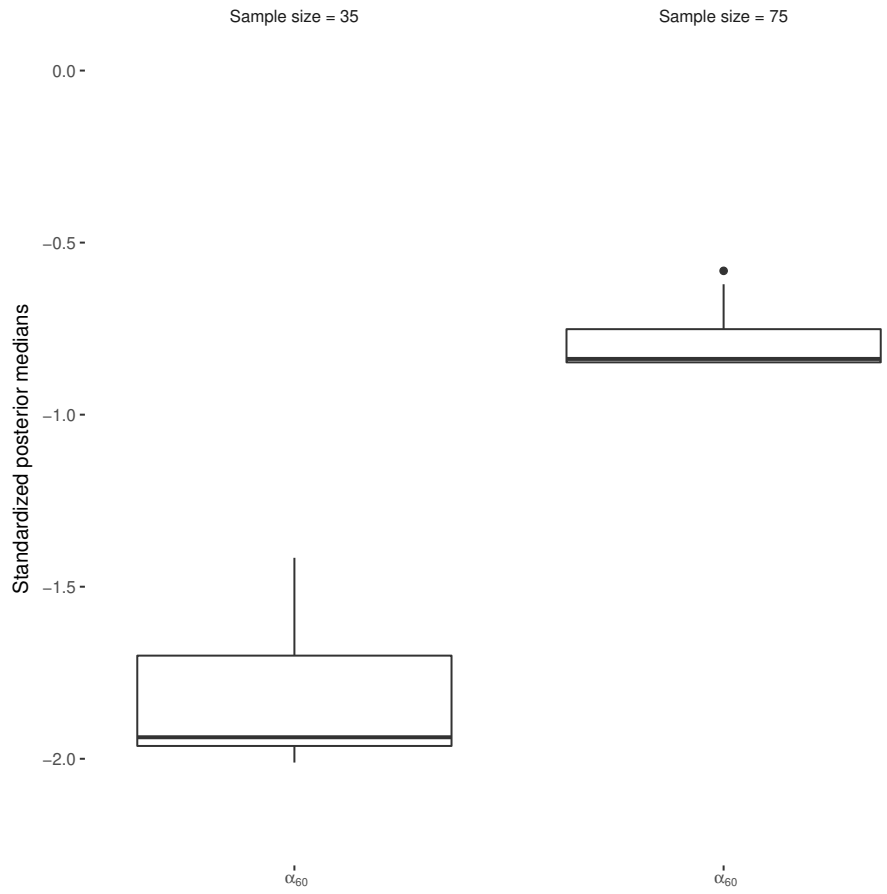

Figure 2.7: Standardized posterior medians for α_{60} in the prior sensitivity analysis.

2.7 Discussion

Bayesian methods are a useful alternative to ML estimation for structural equation models. In the case of small samples, ML estimation can result in empirical weak identification and inadmissible estimates whereas BSEM analyses can prevent these problems. In order to use the BSEM framework, however, prior distributions must be specified for the model parameters. In this paper, we focused on default priors that can be applied in an automatic fashion for a BSEM analysis when prior knowledge is absent or if a researcher does not wish to include external information. Based on the results, we recommend the Mplus default setting (i.e., the noninformative improper prior $\pi(\sigma^2) \propto 1$ for variance parameters, combined with the vague proper prior $\mathrm{N}(0, 10^{10})$ for location parameters) as general default prior for BSEM. In general, we recommend against the use of the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$, since it suffers from major convergence problems, and against the vague proper priors which approximate this improper prior and consequently lead to instable MCMC estimation. The vague proper priors can be considered in the

prior sensitivity analysis, only when multiple values for the hyperparameters are included and the results do not vary across these choices. The performance of the different default priors varied greatly across conditions. For this reason it is highly recommended to consider several default priors when performing a default BSEM analysis, to assess robustness of the results to the choice of the prior.

For $N = 35$, ML estimation performed better than the Bayesian methods in terms of bias. This can be explained by the fact that ML estimation does not force the separate variances to be positive, while the priors we have considered are only defined in the region where the separate variances are positive. In the case of small samples, the likelihood has support for negative variances while the default priors give probability zero to negative variances to obtain interpretable estimates at the cost of introducing some bias. Given that variance parameters are often nuisance parameters, one might argue that minimizing bias is preferred over interpretable estimates. It would be interesting to adopt a Bayesian approach where priors for the variances have support in the negative subspace of certain variance parameters and compare the bias to ML estimation (e.g., Mulder & Fox, 2013). Generally, however, ML estimation cannot be recommended for small samples ($N = 35$) due to the low coverage rates for the variance parameters and high type 1 error rates for testing direct effects. For larger samples (i.e., $N = 75$, 150, 500) ML performed good in terms of all outcome measures and generally outperformed the Bayesian methods. Therefore, if ML estimation is feasible, it is recommended for large samples ($N \geq 75$). Note, however, that for more complex models than the SEM studied here new simulations have to be performed to check whether the sample size of the data at hand is large enough to fit this more complex model using ML. Additionally, ML estimation has the disadvantage that confidence intervals depend on the sampling plan. This implies that optional stopping or deciding to collect more data because of inaccurate results (i.e., wide confidence intervals) is not straightforward to incorporate in a classical analysis using confidence intervals (e.g., Robert, 2007, p. 23). Bayesian credible intervals, on the other hand, abide by the likelihood principle (Berger & Wolpert, 1984), and therefore the results are invariant of the sampling plan.

We have proposed two EB priors which are novel in BSEM. Although the EB priors performed best in terms of MSE, they did not perform well in terms of bias, coverage rates, quantiles, and Type 1 error rates. Several studies have found that EB priors can result in an underestimation of the posterior variance (Carlin & Louis, 2000a; Darnieder, 2011; Efron, 1996), which can partly explain the low coverage rates. Furthermore, the EB priors considered in this paper were developed to be generally applicable. For example, the proposed EB prior for location parameters is centered around zero with the prior variance chosen such that the prior has positive

support where the likelihood is concentrated. However, for some parameters such as intercepts, the prior mean of zero might not be realistic and can lead to biased estimates. In addition, it may be that in certain extreme situations (e.g., when the error variances are approximately zero), data dependent priors, such as our EB priors, should not be used. In general, we believe that the EB methodology offers interesting possibilities for BSEM, however, more research is needed for further development of good EB priors.

We provided guidelines on how to conduct a (default) prior sensitivity analysis in Mplus and illustrated these guidelines on a structural equation model from the literature. An important step is choosing which parameters are of substantive interest. If these parameters are insensitive to the default priors, the estimates can be readily interpreted even if the estimates of some nuisance parameters show prior sensitivity. If the estimated parameters of interest are sensitive to the default priors, i.e., they differ more than the chosen threshold value, the (range of) results of all default priors should be reported. To obtain robust bounds for the interval estimates, we recommend combining the posterior draws of the different default Bayesian analyses, and subsequently, reporting the upper and lower bounds of the 95% credible intervals based on the combined set of draws (e.g., Berger, 2006).

We investigated only conditionally conjugate priors since these are available in Mplus. However, many non-conjugate priors have been proposed in the Bayesian literature as more robust (i.e., less influential) alternatives. For example, Gelman (2006) and Polson and Scott (2012) proposed the half-Cauchy prior for random effects variances, which can be implemented in a Gibbs sampler relatively easy through parameter expansion. A second option for random effects variances is a Gamma prior in combination with posterior mode estimates which has been proposed in the context of meta-analysis by Chung, Rabe-Hesketh, and Choi (2013). Note that the choice of the prior for residual variances is considerably less important than the prior for the variances of latent variables (e.g., Polson & Scott, 2012). For intercept, mean, and regression parameters a robust alternative is the t-distribution, which has been proposed as prior for logistic models by Gelman, Jakulin, Pittau, and Su (2008) and as error distribution to obtain robust models (e.g., robust growth curve models; Zhang, Lai, Lu, & Tong, 2013). The t-distribution includes the Cauchy distribution as special case when the number of degrees of freedom is set to 1. These priors should be investigated in the context of BSEM to assess their performance and determine whether they can be used as default priors.

Appendix: Industrialization and political democracy model in matrix form

The structural model (for $i = 1, \ldots, n$) is given in matrix form as:

$$\begin{pmatrix} \eta_i^{60} \\ \eta_i^{65} \end{pmatrix} = \begin{pmatrix} \alpha^{60} \\ \alpha^{65} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ b_{21} & 0 \end{pmatrix} \begin{pmatrix} \eta_i^{60} \\ \eta_i^{65} \end{pmatrix} + \begin{pmatrix} \gamma^{60} \\ \gamma^{65} \end{pmatrix} \xi_i + \begin{pmatrix} \zeta_i^{60} \\ \zeta_i^{65} \end{pmatrix}$$

With ξ_i representing industrialization level in country i in 1960, and η_i^{60} and η_i^{65} representing political democracy in country i in 1960 and 1965, respectively. The parameters of interest in this model are the direct and indirect effect of industrialization in 1960 on political democracy in 1965, γ_{65} and $\gamma_{60} \cdot b_{21}$, respectively.

The measurement model for political democracy is given by:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \end{pmatrix} = \begin{pmatrix} 0 \\ \nu_2^y \\ \nu_3^y \\ \nu_4^y \\ 0 \\ \nu_6^y \\ \nu_7^y \\ \nu_8^y \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \lambda_2^y & 0 \\ \lambda_3^y & 0 \\ \lambda_4^y & 0 \\ 0 & 1 \\ 0 & \lambda_6^y \\ 0 & \lambda_7^y \\ 0 & \lambda_8^y \end{pmatrix} \begin{pmatrix} \eta_i^{60} \\ \eta_i^{65} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} D_i^{15} \\ D_i^{24} \\ D_i^{26} \\ D_i^{37} \\ D_i^{48} \\ D_i^{68} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i}^y \\ \epsilon_{2i}^y \\ \epsilon_{3i}^y \\ \epsilon_{4i}^y \\ \epsilon_{5i}^y \\ \epsilon_{6i}^y \\ \epsilon_{7i}^y \\ \epsilon_{8i}^y \end{pmatrix}$$

With $\boldsymbol{D}$ representing a vector of pseudo-latent variables used to model the correlations between measurement errors in such a way that the covariance matrix $\boldsymbol{\Sigma_y}$ remains a diagonal matrix.

The measurement model for industrialization level is given by:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix} = \begin{pmatrix} 0 \\ \nu_2^x \\ \nu_3^x \end{pmatrix} + \begin{pmatrix} 1 \\ \lambda_2^x \\ \lambda_3^x \end{pmatrix} \xi_i + \begin{pmatrix} \delta_{1i}^x \\ \delta_{2i}^x \\ \delta_{3i}^x \end{pmatrix}$$

Chapter 3

Bayesian multilevel structural equation modeling: An investigation into robust prior distributions

Based on van Erp, S. and Browne, W.J. (In preparation). Bayesian multilevel structural equation modeling: An investigation into robust prior distributions.

Abstract

Multilevel structural equation modeling (MLSEM) is a popular technique to model latent variables in samples that are grouped in some way. Bayesian estimation of MLSEMs offers specific advantages in terms of sample size requirements and computational feasibility. The latter advantage is especially true when there are categorical indicators, since Markov Chain Monte Carlo sampling does not require multidimensional numerical integration, while existing maximum likelihood methods do. The Bayesian approach, however, does require careful specification of the prior distribution. A well known problem in the Bayesian literature on multilevel models is the specification of the prior for the random effects variance parameter. The traditional "non-informative" conjugate choice of an inverse-Gamma prior with small hyperparameters has been shown time and again to actually be very informative and sensitive to the exact choice of the hyperparameters. As a result, several weakly-informative prior distributions have been proposed as alternative, more robust priors for random effects variances or standard deviations (e.g., the half-cauchy prior). In this paper, we investigate these alternative, state-of-the-art prior distributions in the context of a MLSEM. In contrast to multilevel models without latent variables, MLSEMs have multiple random effects variance parameters, both for the multilevel structure and for the latent variable structure. It is therefore even more important to construct reasonable and robust priors for these parameters.

Keywords: Bayesian, Multilevel, Structural Equation Models, Priors.

3.1 Introduction

Multilevel structural equation modeling (MLSEM) has become increasingly popular to test complex theories in samples that are hierarchically structured (Muthén, 1994). While SEM allows researchers to control for measurement error due to the fact that only a finite number of items are sampled, multilevel analysis takes into account sampling error due to the fact that only a finite number of individuals is sampled. Combined, MLSEM offers a powerful tool that is being used throughout educational, psychological, and sociological research. MLSEMs have also been termed doubly latent multilevel models (DLMM) (Marsh et al., 2009): they include latent variables to account for measurement error and random effects (which are essentially latent variables) to account for the hierarchical structure in the data.

Traditionally, maximum likelihood (ML) algorithms have been used to estimate MLSEMs. However, in the case of categorical indicators, such algorithms require multidimensional numerical integration which quickly becomes computationally infeasible (Asparouhov & Muthén, 2007). Weighted least squares (WLS) algorithms avoid the high dimensional integration, but are restricted to random intercept models (Asparouhov & Muthén, 2012). Additionally, both ML and WLS rely on the frequentist asymptotical framework and thus require a substantial number of groups to obtain admissible and accurate parameter estimates (Hox & Maas, 2001). For example, Meuleman and Billiet (2009) concluded that for relatively simple MLSEMs, at least 60 groups are needed to detect large structural effects at the between level.

Due to the problems with frequentist estimation methods, Bayesian estimation of MLSEMs has become increasingly popular. In a Bayesian analysis, a prior distribution is specified for each of the parameters in the model. Combined with the likelihood of the data this results in a posterior distribution, which is generally obtained through Markov Chain Monte Carlo (MCMC) sampling. The Bayesian approach offers several advantages compared to the frequentist framework. First, MCMC sampling does not require multidimensional numerical integration enabling complex models with categorical indicators to be estimated (see e.g., Asparouhov & Muthén, 2012). Second, estimates of transformed parameters, such as indirect structural effects, can be easily obtained by transforming the MCMC samples and credible intervals are obtained automatically (Y. Yuan & MacKinnon, 2009). Credible intervals are the Bayesian alternative of confidence intervals, but unlike confidence intervals do not rely on normality assumptions or asymptotical theory. Third, through the prior distribution, the problem of variances that are estimated to be negative (i.e., Heywood cases) can be solved; so long as the prior on the variance has zero mass at negative values, the estimate can never become negative. Moreover, prior information about parameters in the model can be included in the prior

distribution. This information might be based on previous investigations or expert knowledge. Finally, it has been shown that Bayesian MLSEM can provide accurate estimates with a smaller number of groups compared to frequentist estimation (Hox et al., 2012; Zitzmann, Lüdtke, Robitzsch, & Marsh, 2016).

A main focus of the Bayesian literature on multilevel models is the specification of the prior for the random effects variance parameter. Historically, inverse-Gamma priors were popular choices for variance parameters, including those at the between level, since they are conjugate and thus result in a posterior that is an inverse-Gamma distribution as well. Specifically, an inverse-Gamma(ϵ, ϵ) prior with small ϵ has often been used as a default uninformative prior distribution. However, it is now well-known in the literature that for random effects variances, this specification is actually very informative and highly sensitive to the exact choice of the hyperparameters (see e.g., Gelman, 2006; Klein & Kneib, 2016; Lunn, Spiegelhalter, Thomas, & Best, 2009). A second often used class of default priors for variance parameters are uniform improper priors. This is in line with the "objective" Bayesian approach (Berger, 2006) since these priors will, in simple models, result in the same estimates as classical ML estimation. Various improper uniform prior options exist. The most straightforward approach is to specify a uniform prior directly on the variance, i.e., $p(\sigma^2) \propto 1$. A second option, advocated by Berger (2006), Berger and Strawderman (1996), and Gelman (2006), is to specify a uniform prior on the standard deviation, i.e., $p(\sigma^2) \propto \sigma^{-1}$. Finally, a uniform prior can be specified on the logarithm of the variance, i.e., $p(\sigma^2) \propto \sigma^{-2}$. The main issue with specifying improper priors for random effects variances is that the resulting posterior might be improper as well in cases where there is a limited amount of information in the data on the higher level (Berger, 2006; Gelman, 2006). This has been shown empirically in a SEM where improper priors resulted in lower convergence rates (van Erp, Mulder, & Oberski, 2018). Note that the inverse-Gamma (ϵ, ϵ) prior approximates the improper prior $p(\sigma^2) \propto \sigma^{-2}$ as the ϵ goes to zero. As a result, the inverse-Gamma prior can also lead to an unstable MCMC sampler.

Recently, several weakly-informative prior distributions have been proposed as alternative priors for random effects variances. For example, Gelman (2006) and Polson and Scott (2012) propose to use the half-Cauchy prior, which is proper and thus avoids the issue of uniform improper priors. Additionally, the half-Cauchy prior has heavy tails and is therefore considered robust to misspecification of the scale. Simpson, Rue, Riebler, Martins, and Sørbye (2017) discuss a general approach to construct so-called "scale dependent" priors, which can be straightforwardly applied to random effects variances. Although some studies exist that compare some of these priors in general multilevel models (see e.g., Klein & Kneib, 2016), they have yet to be investigated in the context of MLSEM. Studies investigating prior choice

in MLSEM generally focus on conjugate prior distributions that vary in the degree of informativeness (Depaoli & Clifton, 2015; Helm, 2018; Zitzmann et al., 2016). These studies show that informative priors, when correctly specified, perform best. However, in reality we do not know the true population value and therefore require prior distributions that are robust against possible prior misspecification. Moreover, MLSEMs have multiple random effects variance parameters, both for the multilevel structure and for the latent variable structure. It is therefore even more important to construct reasonable and robust priors for these parameters.

The goal of this study is to compare robust prior distributions for random effects variances in MLSEM. We will compare default specifications of the robust priors as well as informative (in)accurate specifications. The outline of this paper is as follows: in Section 3.2 we descibe an empirical application that will be used to illustrate the MLSEM, followed by a description of the model in Section 3.3. The robust prior distributions are presented and compared in Section 3.4 and applied to the empirical application in Section 3.5. Section 3.6 presents three simulation studies to investigate the performance of the priors, followed by a discussion in Section 3.7.

3.2 Empirical application

To illustrate the influence of the various prior distributions, we will focus on a specific application of multilevel SEM to estimate the so-called "Big-Fish-Little-Pond-Effect" (BFLPE). The BFLPE is a contextual effect that relates to students' achievement and academic self-concept. Academic self-concept refers to students' perception of their academic abilities and competencies. The BFLPE predicts that at the individual (or within) level, student achievement has a positive effect on academic self-concept. However, at the between level, school-average achievement has a negative effect on academic self-concept. Thus, students with equal levels of achievement are expected to have a higher academic self-concept when they are in low-ability classes or schools compared to students in high-ability classes or schools. This has important policy implications, especially given the many correlations between academic self-concept and future academic performance (see e.g., Guay, Marsh, & Boivin, 2003; Marsh & Craven, 2006).

A large body of research has consistently found this effect to exist across countries, age groups, and academic domains (see for example the overview by Marsh et al., 2008, and the references therein). Here, we replicate the analysis by Nagengast and Marsh (2011) who use data from the 2006 round of PISA (OECD, 2007) to investigate the BFLPE with regard to science in the United Kingdom. For the purpose of illustration, we will only reanalyse the data from 2444 students in 98 Scottish schools. Science achievement is measured using various open- and

closed-format problems and test scores for each student are reported by five plausible values. Plausible values are random draws from the posterior distributions based on each primary test score and are used in PISA to increase the measurement accuracy. Every analysis is run separately for each plausible value and the results are combined appropriately to obtain the final estimates[1]. Science academic self-concept was measured using six items with a 4-point Likert scale, with higher values corresponding to more positive self-concept. Although the selection probabilities of students in PISA vary, the use of survey weights is not straightforward in Bayesian analysis. Since the aim of this application is to illustrate the influence of the different priors, we do not include survey weights in the analysis. Moreover, we remove all cases with missing data, resulting in a total of 2238 observations. We will now turn to a discussion of the multilevel SEM used to estimate the BFLPE.

3.3 Bayesian doubly latent ordinal multilevel model

Model

To estimate contextual effects like the BFLPE, Marsh et al. (2009) proposed the use of doubly latent multilevel models (DLMM). The term "doubly latent" arises from the fact that the model takes into account both measurement and sampling error. Measurement error, which is a consequence of the fact that only a finite number of items are sampled is controlled for by specifying a measurement model for each of the factors. Sampling error, on the other hand, is the result of sampling only a finite number of individuals and is controlled by including random effects in the model.

Note that in the case of the PISA data, science academic achievement is measured by one indicator (i.e., the plausible value), such that it is not possible to include a measurement model for this variable. Consequently, the model considered in this paper is doubly latent for the dependent factor academic self-concept, but only accounts for sampling error, and not for measurement error, for academic achievement.

The items for academic self-concept are measured on a 4-point Likert scale for individual i in group j measured on item k. To model such ordinal variables, let us assume that the observed responses y_{ikj}, which take on the values $\{1, 2, \ldots, C_k\}$

[1]With a Bayesian analysis, results can be combined either by using Rubin's rules on the posterior estimates or by combining the posterior draws across analyses and computing the posterior summaries based on the combined draws. Here, we used the latter approach which is recommended when posterior densities deviate from normality (Zhou & Reiter, 2010).

where C_k denotes the number of categories for item k, are generated by a latent continuous variable $\tilde{y}_{ikj}$ as follows:

$$\{y_{ikj} = c_k\} \Leftrightarrow \{\gamma_{c_k-1,k} < \tilde{y}_{ikj} \leq \gamma_{c_k,k}\} \tag{3.1}$$

for $c_k = 1, \ldots, C_k$ categories. In our case, $C_k = 4$ for all items k. The measurement model can then be defined for the continuous latent responses $\tilde{y}_{ikj}$ at the within level:

$$\tilde{y}_{ikj} = \mu_{kj} + \lambda_k^W \eta_{ij}^W + \epsilon_{ikj}^W, \tag{3.2}$$

where μ_{kj} denotes the intercept for item k in group j, λ_k^W is the loading for item k at the within level, η_{ij}^W is the factor score for individual i in group j, and ϵ_{ikj}^W is the residual at the within level for individual i in group j measured on item k. Throughout this paper, we assume a logistic distribution for the residuals at the within level, i.e., $\epsilon_{ikj}^W \sim \text{logistic}(0, \sigma_{W,k}^2)$. We focus on the random intercept model, such that the measurement model at the between level is defined as:

$$\mu_{kj} = \mu_k + \lambda_k^B \eta_j^B + \epsilon_{kj}^B, \tag{3.3}$$

where μ_k reflects the overall intercept for item k, λ_k^B is the loading for item k at the between level, η_j^B is the factor score at the between level for group j, and ϵ_{kj}^B denotes the residual at the between level for item k in group j, which we assume to be normally distributed, i.e., $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$. Combining the measurement model at the within and between level, we obtain:

$$y_{ikj} = \mu_k + \lambda_k^W \eta_{ij}^W + \epsilon_{ikj}^W + \lambda_k^B \eta_j^B + \epsilon_{kj}^B \tag{3.4}$$

The factor scores on self-concept are predicted by the academic achievement scores x_{ij} leading to the following structural model:

$$\eta_{ij}^W \sim N(\beta^W x_{ij}, \omega_W^2) \tag{3.5}$$

$$\eta_j^B \sim N(\alpha + \beta^B x_{b\cdot j}, \omega_B^2) \tag{3.6}$$

Here, β^W represents the effect of achievement on self-concept at the within level, whereas β^B represents the effect of achievement on self-concept at the between level.

As recommended by Asparouhov and Muthén (2019), we use latent mean centering which estimates the group mean achievement in each group j, i.e., $x_{b\cdot j}$ to take into account measurement error. The parameter of interest is the contextual effect, which equals $\beta^B - \beta^W$.

An important quantity in multilevel models is the variance partition coefficient (VPC; Goldstein, Browne, & Rasbash, 2002), which is a measure of the correlation between students in the same school. In multilevel SEM, the VPC ρ_k for item k is defined as (see e.g., Muthén, 1994):

$$\rho_k = \frac{\lambda_k^{2,B}\omega_B^2 + \sigma_{B,k}^2}{(\lambda_k^{2,B}\omega_B^2 + \sigma_{B,k}^2) + (\lambda_k^{2,W}\omega_W^2 + \sigma_{W,k}^2)} \tag{3.7}$$

Identification restrictions

Certain restrictions are needed to identify the model. Specifically, we need to identify the location and scales of the latent variables in the model. First, consider the ordinal measurement model at the within level. Here, we are assuming $\tilde{y}_{ikj}$ to be generated by an underlying continuous latent variable. To identify this part of the model, we fix the mean of the latent response to zero by setting μ_k to zero for each item k and we fix the variance to that of the standard logistic distribution, i.e., $\sigma_{W,k}^2 = \pi^2/3$. Alternatively, a probit link function can be used, in which case $\epsilon_{ikj}^W \sim N(0,1)$. Second, we need to identify the latent variable η^W and η^B at the within and between level. Again, we fix the mean to zero by restricting $\alpha = 0$. To identify the latent scales we set one loading to 1 at both levels, i.e., $\lambda_1^B = \lambda_1^W = 1$. These restrictions are, in theory, sufficient to identify the model. However, it is generally recommended to additionally impose cross-level invariance, i.e., $\boldsymbol{\lambda^W = \lambda^B}$ to improve interpretability of the factors at both levels and to avoid estimation issues (Jak, 2018).

Prior distributions

In a Bayesian analysis, prior distributions need to be specified for each parameter in the model, i.e., $p(\gamma_{c_k,k}, \lambda_k, \sigma_{B,k}^2, \beta^W, \omega_W^2, \beta^B, \omega_B^2)$. The focus of this paper is on priors for the random effects variances $\sigma_{B,k}^2, \omega_W^2$, and ω_B^2 and the next section discusses the priors we consider for these parameters in detail. For the other parameters in the model, we specify the following independent, uninformative or weakly informative prior distributions:

$$
\begin{aligned}
p(\gamma_{c_k,k}) &\propto 1, \text{ for } k = 1, \ldots, 6 \\
p(\lambda_k) &\sim \text{Normal } (0, 5), \text{ for } k = 1, \ldots, 6 \\
p(\beta^W) &\sim \text{Normal } (0, 1) \\
p(\beta^B) &\sim \text{Normal } (0, 1).
\end{aligned}
$$

Note that other choices are possible and it is always important to investigate the sensitivity of the results to the specification of the prior distribution. We now turn to a discussion of the priors for the random effects variance parameters.

3.4 Robust prior distributions for random effects variances

The problems with the traditional inverse-Gamma family and the improper uniform priors for random effects variances have inspired many researchers to investigate alternative, more robust prior distributions. Here, we focus on three such alternative classes of priors: 1) Student's t priors; 2) F priors; and 3) Scale dependent priors. Note that some priors are specified on the variances while others are specified on the standard deviations. Throughout this section, we will use θ to denote standard deviations and θ^2 to denote variance parameters.

Student's t family

A widespread proposal for the prior on random effects is to specify a proper, weakly-informative half-t distribution on the standard deviation (see for example, Gelman, 2006; Polson & Scott, 2012). The probability density function of the half-t distribution is given by:

$$
p(\theta; \nu, \sigma_0^2) = \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma_0^2}} \left(1 + \frac{\theta^2}{\nu\sigma_0^2}\right)^{-\frac{\nu+1}{2}}, \tag{3.8}
$$

where σ_0^2 denotes the scale and ν the degrees of freedom. Different choices for the scale parameter σ_0^2 are discussed in Subsection 3.4. For the degrees of freedom ν, smaller values result in heavier tails. For $\nu = \infty$, we obtain the half-normal distribution, which has been proposed, among others, by Frühwirth-Schnatter and Wagner (2010). Roos, Held, et al. (2011) have shown that the half-normal prior indeed results in estimates of the random effects variances that are less sensitive to the chosen hyperparameters compared to the inverse-Gamma prior. Nevertheless,

the half-normal distribution still has very light tails and will therefore be less robust to variance parameters that are larger than expected. Therefore, a more robust proposal in the literature is to specify the degrees of freedom to be smaller. Gelman (2006) notes the special case of the half-Cauchy prior with $\nu = 1$ as a reasonable default option. The half-Cauchy prior has been further investigated by Polson and Scott (2012), who have shown that it has excellent frequentist risk properties. There are some reports, however, that the heavy tails of the half-Cauchy prior can lead to numerical difficulties in MCMC sampling (Ghosh, Li, & Mitra, 2018; Piironen & Vehtari, 2015). This issue generally arises in situations when there is not enough information in the data such that the parameter is weakly or non-identified. The heavy tails of the (half-)Cauchy prior do not perform enough shrinkage in such cases to identify the parameter.

F family

A popular family of distributions for variance parameters in the Bayesian literature is the family of F priors, also known as scaled Beta2 or generalized beta prime priors (Fúquene, Pérez, & Pericchi, 2014; Mulder & Pericchi, 2018; Pérez, Pericchi, & Ramírez, 2017; Polson & Scott, 2012). The probability density function of the F prior is given by (Mulder & Pericchi, 2018)[2]:

$$p(\theta^2; \nu_1, \nu_2, \sigma_0^2) = \frac{\Gamma(\frac{\nu_2+\nu_1}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(\sigma_0^2)^{\frac{\nu_1}{2}}}(\theta^2)^{\frac{\nu_1}{2}-1}(1+\theta^2/\sigma_0^2)^{-\frac{\nu_1+\nu_2}{2}}, \tag{3.9}$$

where $\sigma_0^2 > 0$ denotes a scale parameter, the degrees of freedom $\nu_1 > 0$ determines the behaviour around zero, and the degrees of freedom $\nu_2 > 0$ influences the tail behaviour in a similar manner as the degrees of freedom for Student's t distribution. Specifically, $\nu_2 = 2$ results in similar tails as a Cauchy distribution. With regard to the behaviour at zero, $\frac{\nu_1}{2} > 1$ results in a density that is zero at the origin, for $\frac{\nu_1}{2} = 1$ the density is bounded at the origin, and for $\frac{\nu_1}{2} < 1$ the density goes to infinity at the origin.

Interestingly, if $p(\theta^2) \sim \text{F}(\nu_1 = 1, \nu_2, \sigma_0^2)$, then the prior on the standard deviation $p(\theta)$ corresponds to a half-Student's t distribution with degrees of freedom equal to ν_2 and scale $\sqrt{\sigma_0^2}$. Thus, the F family can be seen as a generalization of the Student's t family. Moreover, if $p(\theta^2) \sim \text{F}(\nu_1, \nu_2, \sigma_0^2)$, then the prior on the precision $h = \frac{1}{\theta^2}$ will also be an F distribution, with a scale of $\frac{1}{\sigma_0^2}$ (i.e., the reciprocality property).

[2] This corresponds to the parametrization used in Pérez et al. (2017) when $p = \frac{\nu_1}{2}$, $q = \frac{\nu_2}{2}$, and $b = \sigma_0^2$

To implement the F prior, the following Gamma mixture of inverse-Gamma parametrization can be used:

$$p(\theta^2) \sim \text{IG}(\frac{\nu_2}{2}, \tau^2) \tag{3.10}$$

$$p(\tau^2) \sim \text{G}(\frac{\nu_1}{2}, \sigma_0^{-2}), \tag{3.11}$$

which results in the F prior when integrating out τ^2.

Scale dependent family

Simpson et al. (2017) have proposed the use of penalised complexity priors as general default priors. The basic idea of these priors is that deviations from a simpler base model are penalised thereby avoiding a model that overfits. This characteristic is especially useful in the context of random effects variance parameters which are generally weakly identified due to the small amount of information available at the higher level. Simpson et al. (2017) derive the penalised complexity prior for the precision of a random effect as follows: first, define the base model. For example, for the random effects component $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$ in 3.3, the base model, which is the simplest model in its class, would correspond to $\sigma_{B,k}^2 = 0$ or the absence of random effects. Next, the distance of the flexible extension of the base model, i.e., $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$ to the base model is computed based on the Kullback-Leibler divergence (KLD). Simpson et al. (2017) define the distance between the two models with densities f and g as: $d(f||g) = \sqrt{(2KLD(f||g))}$. Then a prior is specified for the distance d such that deviations from the base model are penalised. By assuming a constant rate of penalisation, Simpson et al. (2017) specify an exponential prior for the distance $p(d) = \lambda \exp(-\lambda d)$, although they note that it is possible to relax this assumption if needed. For example, when interest lies in variable selection, the exponential tails are too light and a heavier tailed prior can be more sensible, for example a half-Cauchy prior on the distance will recover the well-known horseshoe prior. Transforming the prior on the distance to the original space leads to a type-2 Gumbel prior for the precision or, equivalently, an exponential distribution with scale σ_0^2 for the standard deviation[3]:

$$p(\theta; \sigma_0^2) = \frac{1}{\sigma_0^2} \exp(-\frac{1}{\sigma_0^2}\theta) \tag{3.12}$$

[3]We parametrise the exponential distribution in terms of the scale σ_0^2 for consistency, but generally it is defined using the rate parameter λ, which is the inverse of the scale, i.e., $\lambda = \frac{1}{\sigma_0^2}$.

Lower values for σ_0^2 will result in an increased penalisation for deviating from the base model. The choice of σ_0^2 will be discussed in Subsection 3.4.

Specification of the hyperparameters

Table 3.1: Overview of robust prior distributions for random effects variances

Prior	Parameter	Scale	Tail behaviour	Behaviour origin
half-Normal	θ	σ_0^2	$\nu = \infty$	-
half-Student's t	θ	σ_0^2	ν	-
half-Cauchy	θ	σ_0^2	$\nu = 1$	-
F	θ^2	σ_0^2	ν_2	ν_1
Exponential	θ	σ_0^2	-	-

Note. θ denotes a standard deviation parameter and θ^2 denotes a variance parameter.

An overview of the different priors considered is provided in Table 3.1. It is clear from Table 3.1 that the priors vary in flexibility. Each prior has a parameter that determines the scale, or how spread out the prior is. In addition, the half-Student's t and F priors have a parameter that influence the tail behaviour of these priors. Specifically, this degrees of freedom hyperparameter determines how heavy the tails are and thus how much prior mass is put on extreme values. In general, a prior with heavier tails is more robust since extreme values will not be shrunken towards zero as much compared to a prior with thinner tails. For the half-Student's t prior, we consider the special cases of the half-normal prior ($\nu = \infty$) and the half-Cauchy prior ($\nu = 1$), but other values for ν can be used, with higher degrees of freedom leading to thinner tails. For the F prior, we will consider $\nu_2 = 0.5$ throughout this paper to obtain tails that are heavier than the Cauchy. The larger ν_2, the thinner the tails will be. Additionally, the F prior has a parameter that determines the behaviour at the origin. This degrees of freedom hyperparameter, ν_1, is set to 1 so that the density goes to infinity at zero. Although other values for ν_1 are possible, it is important to ensure that $\frac{\nu_1}{2} \leq 1$ so that the prior has support at zero.

Throughout this paper, we consider and compare various settings for the scales in each prior (σ_0^2). A general informal recommendation for a default prior is a half-Student's t prior with scale equal to 1 (see, for example, Stan Development Team, 2015, the section on priors for scale parameters in hierarchical models). Therefore, we use $\sigma_0^2 = 1$ throughout this paper as default prior specification. Additionally, if prior information is available, this can be included in the priors as follows. Specify

a value for the standard deviation C and a threshold α such that the probability $P(\theta > C) = \alpha$. This equation can then be solved for θ using the cumulative distribution function. Section 3.6 describes which values for C and α we will consider in the simulation study and provide some examples of the resulting prior scales θ.

Comparison of the priors

In this section, we compare the probability density functions of the various priors to better understand their behaviour. Figure 3.1 shows the densities for the standard deviation. We use a scale of 1 for the robust priors and hyperparameters equal to 0.1 for the inverse-Gamma prior. The main difference between the inverse-Gamma prior and the more robust alternatives is that the inverse-Gamma prior has zero density at zero. As a result, the prior favors standard deviations that are away from zero which is problematic especially for standard deviations of random effects which are often very close to zero. The other priors do have prior probability around standard deviations equal to zero and are therefore more in line with what we might expect in reality.

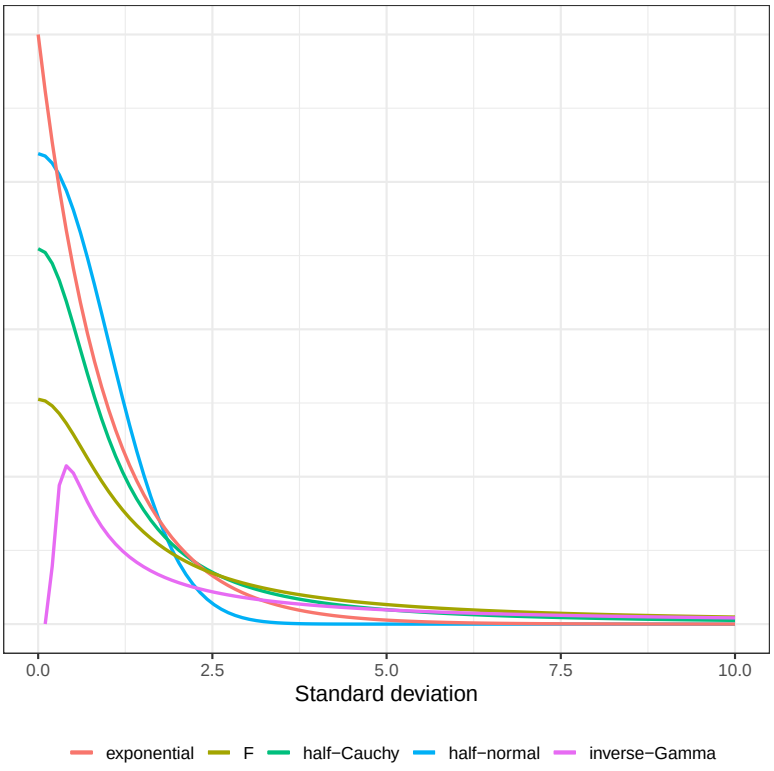

Figure 3.1: Densities of the priors on the standard deviations.

There are multiple standard deviations in the model for which a prior needs to be specified, specifically: $\sigma_{B,k}$, ω_W, and ω_B. Together with the factor loadings at the between and within level, λ_k^B and λ_k^W, these standard deviations play a role in

computing the VPC ρ_k for each item (see equation (3.7)). As a result, specifying a prior on the standard deviations and factor loadings implies a certain prior on the VPC. This prior is considered in Figure 3.2. Recall from Section 3.3 that we fixed the loading of one item to 1 for identification purposes. The left figure shows the VPC of the restricted item while the figure on the right shows the VPC of the unrestricted items. It is clear that this restriction has consequences for the implied prior of the VPC. This is due to the fact that the VPC depends on the loadings (see equation (3.7)). Specifically, we see that some priors put less probability on a VPC close to 1 for the restricted item compared to the free items. This is most pronounced for the priors with the thinnest tails, i.e., the half-normal and exponential priors. Furthermore, the implied priors on the VPC are not symmetrical. This is due to the fact that we assume a standard logistic distribution for the residuals at the within level. As a result, $\sigma^2_{W,k}$, which arises in the denominator of equation (3.7) is fixed as well.

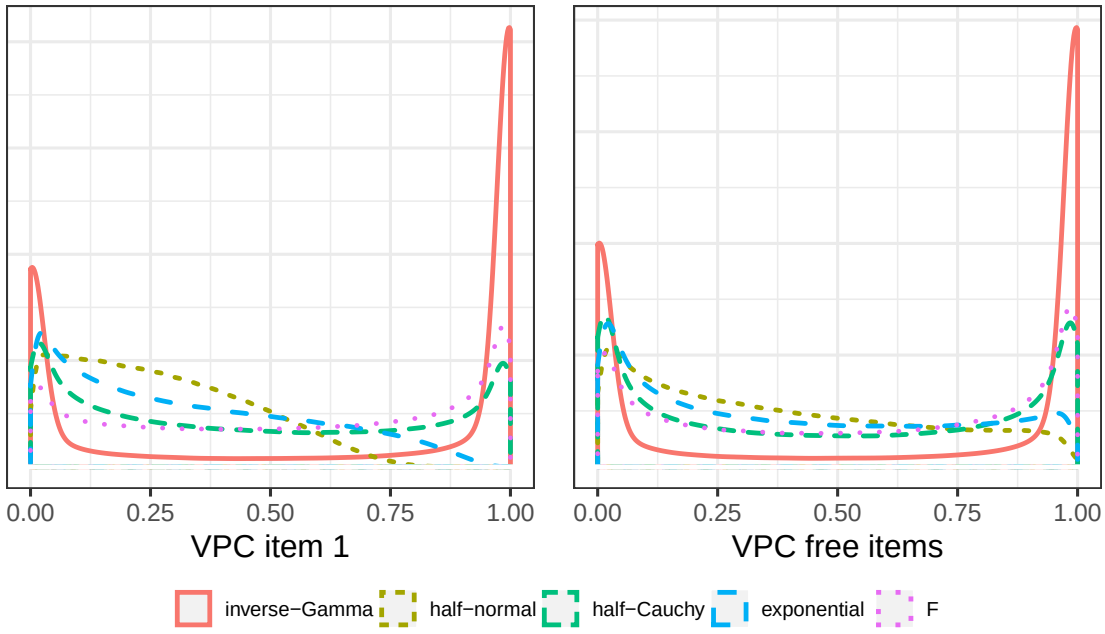

Figure 3.2: Densities of the implied priors on the variance partition coefficient (VPC) of the restricted and free items.

3.5 Priors applied to the empirical example

In this section, we will illustrate the various prior distributions on the empirical data described in Section 3.2. Specifically, we will use the default prior setting with $\sigma_0^2 = 1$. We analyse the data from 2238 students in 98 Scottish schools (excluding all cases with missing data) and we also consider a subset of 233 students in 10 of the schools. We can expect the prior to be more influential in the latter case, since there

is less data to dominate the posterior. One problem with the model using latent group mean centering is that it resulted in multiple analyses that did not converge. Recall that the PISA data uses five plausible values for science achievement, resulting in five analyses for each prior. However, if one of these analyses did not converge, the combined results across plausible values cannot be trusted. For the application, we therefore relied on observed group mean centering instead of latent mean centering. Here, we only report the results for selected parameters of interest. The full results are available online at https://osf.io/pq8gm/.

Figure 3.3 shows the posterior mean estimates and 95% credible intervals for ω_B for the full (right) and partial (left) data set. The results are quite similar across the robust and uniform priors. As expected, the results differ for the inverse-Gamma priors and the estimates are sensitive to the exact choice of the hyperparameters for the inverse-Gamma prior. The differences are less pronounced for the full data set compared to the partial data set. Although not shown, the results for the residual variances at the between level, σ_Y^B show a similar pattern: no substantial differences between the robust and uniform priors, but substantial differences for the inverse-Gamma priors.

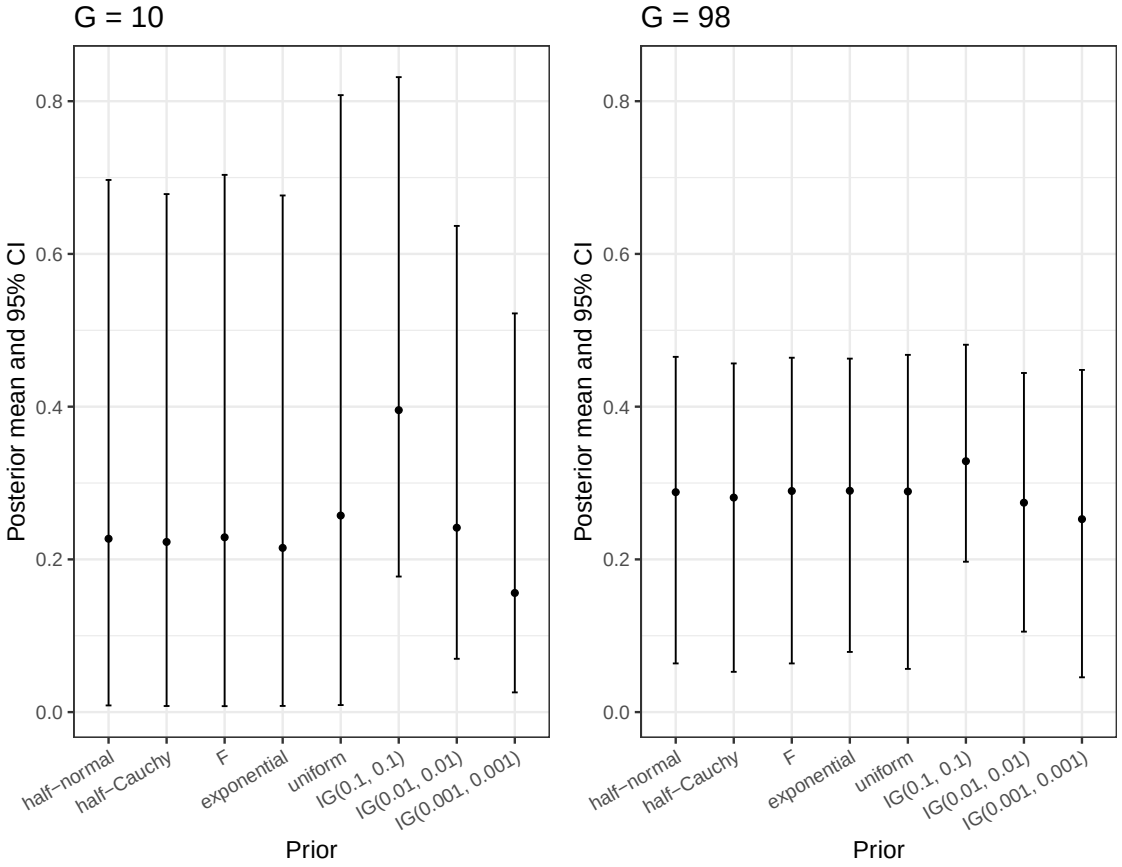

Figure 3.3: Posterior mean estimates and 95% credible intervals for the latent variable standard deviation at the between level ω_B for the full ($G = 98$) and partial ($G = 10$) data.

Figure 3.4 shows the posterior mean estimates and 95% credible intervals for ω_W for the full (right) and partial (left) data set. For the partial data set, we see some slight differences between the priors, with priors with thinner tails (such as the half-Normal) pulling the standard deviation slightly more towards zero compared to the heavier-tailed or uniform priors. However, the differences are small and we do not see substantially different results for the inverse-Gamma prior. Within-level variances are generally less sensitive to the prior because there is more information in the data about these parameters to inform the posterior, compared to between-level variances.

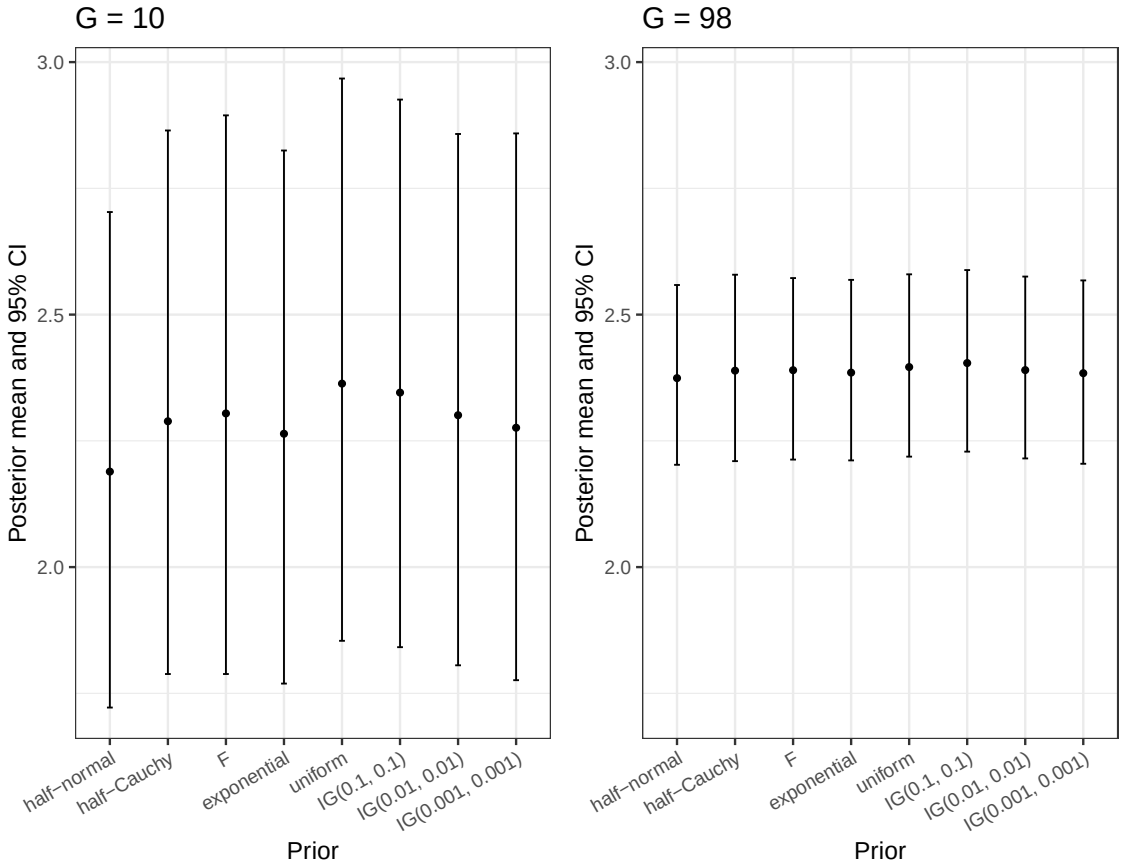

Figure 3.4: Posterior mean estimates and 95% credible intervals for the latent variable standard deviation at the within level ω_W for the full ($G = 98$) and partial ($G = 10$) data.

Ultimately, we are interested in the contextual effect, $\beta_B - \beta_W$, which in this application corresponds to the BFLPE. Note that the BFLPE implies a negative contextual effect. Figure 3.5 shows the posterior densities for the contextual effect for the full (right) and partial (left) data set. Despite the varying estimates for the standard deviations, especially for $G = 10$, the influence of the prior on the contextual effect is negligible. Each prior results in a posterior mean estimate for the BFLPE of approximately -0.004. We can thus conclude that there is no evidence

for a BFLPE in the Scottish schools.

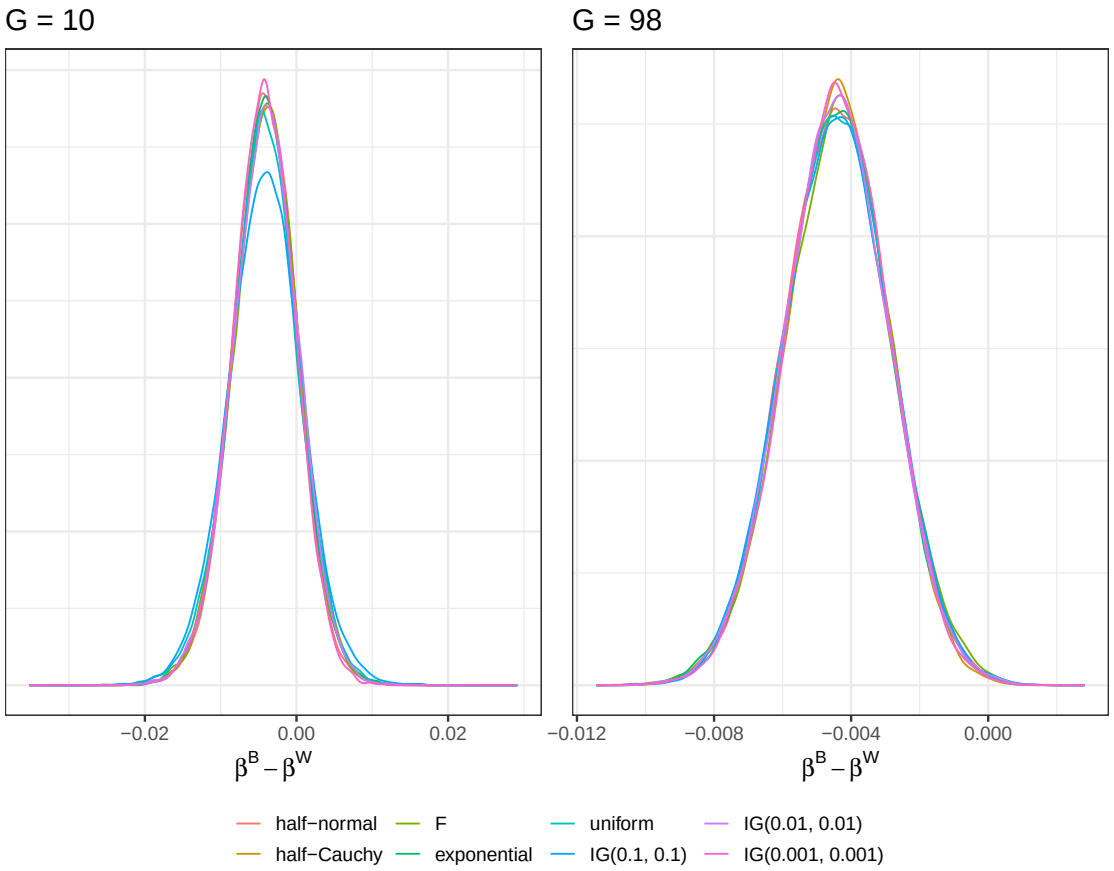

Figure 3.5: Posterior densities for the contextual effect $\beta_B - \beta_W$ for the full ($G = 98$) and partial ($G = 10$) data.

3.6 Simulation studies

In order to investigate the performance of the robust priors for the random effects variances, we conduct three simulation studies. In the first study, we vary the population values for various parameters while keeping the number of groups fixed. We focus on comparing the traditional priors for variance parameters to a default setting of the more robust alternatives. In the second study, we use the same generated data sets as in the first study, but now we focus on informative prior settings with either correct or incorrect information to determine how robust the priors are to misspecifications. Finally, since the influence of the prior distribution decreases as the sample size increases, we vary the number of groups in the third study for a subset of the population values from study 1. We analyse the data sets in study 3 with the default and correct informative prior settings. Below, we describe the conditions and rationale behind them in more detail per study.

Study 1: Influence of population values

Since the focus of this study is on priors for random effects variances, the population values for these parameters are the most important quantities to vary. As a second variable, we consider various effect sizes for the contextual effect, which is generally the parameter of interest. Specifically, we consider the following population values:

1. ***Population values for the variance parameters.*** The sensitivity of the inverse Gamma priors to the hyperparameter choice arises especially in situations in which the random effects variance is close to zero (Gelman, 2006). On the other hand, it is of interest to consider variance parameters that are large compared to the default prior specification in which the scale parameter equals 1. This gives an indication of the robustness of each of the priors. Therefore, we consider the following population values for the latent variable variance parameters: $\omega^2 = c(.01, 1)$. Depending on which value is specified for each variance parameter, the VPC will vary. Hedges and Hedberg (2007) provide an overview of VPC values for educational achievement based on longitudinal surveys conducted in the United States. They found VPCs ranging from .03 to .3. These values are in line with VPCs used in other simulation studies of multilevel SEMs (Depaoli & Clifton, 2015; Helm, 2018; Zitzmann et al., 2016). Therefore, we consider several combinations of the population values for the variance parameters that result in VPCs within this range. Note that the VPC depends on the following parameters: VPC $= f(\lambda_k^{2,B}, \omega_B^2, \sigma_{B,k}^2, \lambda_k^{2,W}, \omega_W^2, \sigma_{W,k}^2)$ (see (3.7)). We fix the population values for $\lambda_k^{2,B}$ and $\lambda_k^{2,W}$ to 1, and $\sigma_{W,k}^2$ to $\frac{\pi^2}{3}$ for all items. We consider different combinations of population values for the variances of the latent variables, ω_B^2 and ω_W^2 and subsequently compute $\sigma_{B,k}^2$ given these values and a VPC $= c(.03, .3)$. An overview of the different population values for the variance parameters and the resulting VPC is presented in Table 3.2.

2. ***Population values for the contextual effect.*** When using group-mean centering, the contextual effect is defined as $\beta^B - \beta^W$. Marsh et al. (2009) provide three different standardised effect size measures that can be interpreted similarly to Cohen's d. They recommend to use either one of the two more conservative effect sizes. Here, we rely on the second effect size in Marsh et al. (2009) to define the population values for the regression coefficients at the between and within level. This effect size standardises the contextual effect with respect to the total variance of self-concept at the within level, i.e., η_W, and is defined as:

$$ES = \frac{2\beta^C \sigma_x^B}{\sigma_x^W (\beta^W)^2 + \omega^W}, \tag{3.13}$$

where β^C denotes the contextual effects which is equal to $\beta^B - \beta^W$ in the case of group-mean centering, and σ_x^B and σ_x^W denote the standard deviations of the predictor at the between and within level, respectively. We consider a small negative standardised contextual effect and no contextual effect, i.e., $ES = c(-0.2, 0)$. We choose these values such that we are able to investigate the power to detect a small contextual effect as well as the type 1 error rate. We fix β^W to 0.2 and σ_x^B and σ_x^W to 1 and compute the value for β^B to obtain the required standardised effect size. The resulting values are shown in Table 3.2.

Table 3.2: Overview population values for the variance parameters and variance partition coefficient.

Setting	ω_B^2	ω_W^2	$\sigma_{B,k}^2$	VPC	ES	β^B
1	0.01	0.01	0.11	0.03	0.00	0.20
2	0.01	0.01	1.42	0.30	0.00	0.20
3	1.00	0.01	1.10	0.03	0.00	0.20
4	1.00	0.01	2.41	0.30	0.00	0.20
5	0.01	1.00	0.14	0.03	0.00	0.20
6	0.01	1.00	1.85	0.30	0.00	0.20
7	1.00	1.00	1.13	0.03	0.00	0.20
8	1.00	1.00	2.84	0.30	0.00	0.20
9	0.01	0.01	0.11	0.03	-0.20	0.19
10	0.01	0.01	1.42	0.30	-0.20	0.19
11	1.00	0.01	1.10	0.03	-0.20	0.19
12	1.00	0.01	2.41	0.30	-0.20	0.19
13	0.01	1.00	0.14	0.03	-0.20	0.10
14	0.01	1.00	1.85	0.30	-0.20	0.10
15	1.00	1.00	1.13	0.03	-0.20	0.10
16	1.00	1.00	2.84	0.30	-0.20	0.10

Combinations of the various population values result in 16 different conditions (2 values ω_B^2 $\times$ 2 values ω_W^2 $\times$ 2 values VPC $\times$ 2 values ES). In this first study, we consider these 16 conditions for a balanced design with 20 groups and a sample

size of 20 within each group. We use 20 groups based on Hox et al. (2012) who concluded that 20 groups are sufficient for multilevel SEM. Moreover, 20 groups is still feasible in terms of data collection. The sample size of 20 within each group is based on educational contexts in which this can be seen as a realistic size for a school class.

We generate the data based on a doubly latent multilevel model such as the one in Section 3.3. The only difference with the model in Section 3.3 is that we use binary instead of ordinal indicators to keep the computation time feasible. We analyse the data sets using the traditional uniform prior and the inverse-Gamma prior with $\epsilon = .1, .01, .001$ as well as the robust half-Normal, half-Cauchy, F, and Exponential priors. For the robust priors we consider a default setting in which the scale equals $\sigma_0^2 = 1$, following a general recommendation often made for a robust default specification (see also Subsection 3.4). This leads to a total of 8 different prior distributions. For the other parameters in the model, we use the same weakly informative priors as in Section 3.3.

Study 2: Influence prior misspecifications

In the second study, we use the same data sets as in Study 1 but now we focus on informative prior distributions. Specifically, we choose the prior scale σ_0 in such a way that $P(\sigma_0 > \sigma_{true}) = \alpha$, where σ_{true} equals the population value under which that parameter was simulated. We consider two settings: in the "correct" informative prior, $\alpha = 0.5$ such that the resulting prior has sufficient probability around the population value. In the "incorrect" informative prior, on the other hand, $\alpha = 0.05$ such that the prior has only very little mass on the population value. Since we never know the true value in practice, it is important to consider this situation to assess the robustness of the various priors to possible misspecifications. In total, we consider 8 different prior specifications (4 types of priors $\times$ 2 settings) for 16 population conditions in Study 2. Figures 3.6 and 3.7 show the various informative prior densities used in the simulation study when the population value for the standard deviation equals 0.1 or 1, respectively. Note that the incorrect F prior specification is missing when the population standard deviation equals 0.1, since this setting resulted in a scale σ_0^2 of 0, whereas the scale should be positive. However, in order to be consistent across priors in specifying the incorrect setting, we did not consider an alternative specification that did result in a positive scale. All priors are more peaked around zero in the incorrect specification, and therefore have less prior mass around the true population value compared to the correct specification. It might appear that both the half-Cauchy and F priors have almost no prior mass around the true population value, but this is just a visual consequence of these priors being very spread out due to their heavy tails.

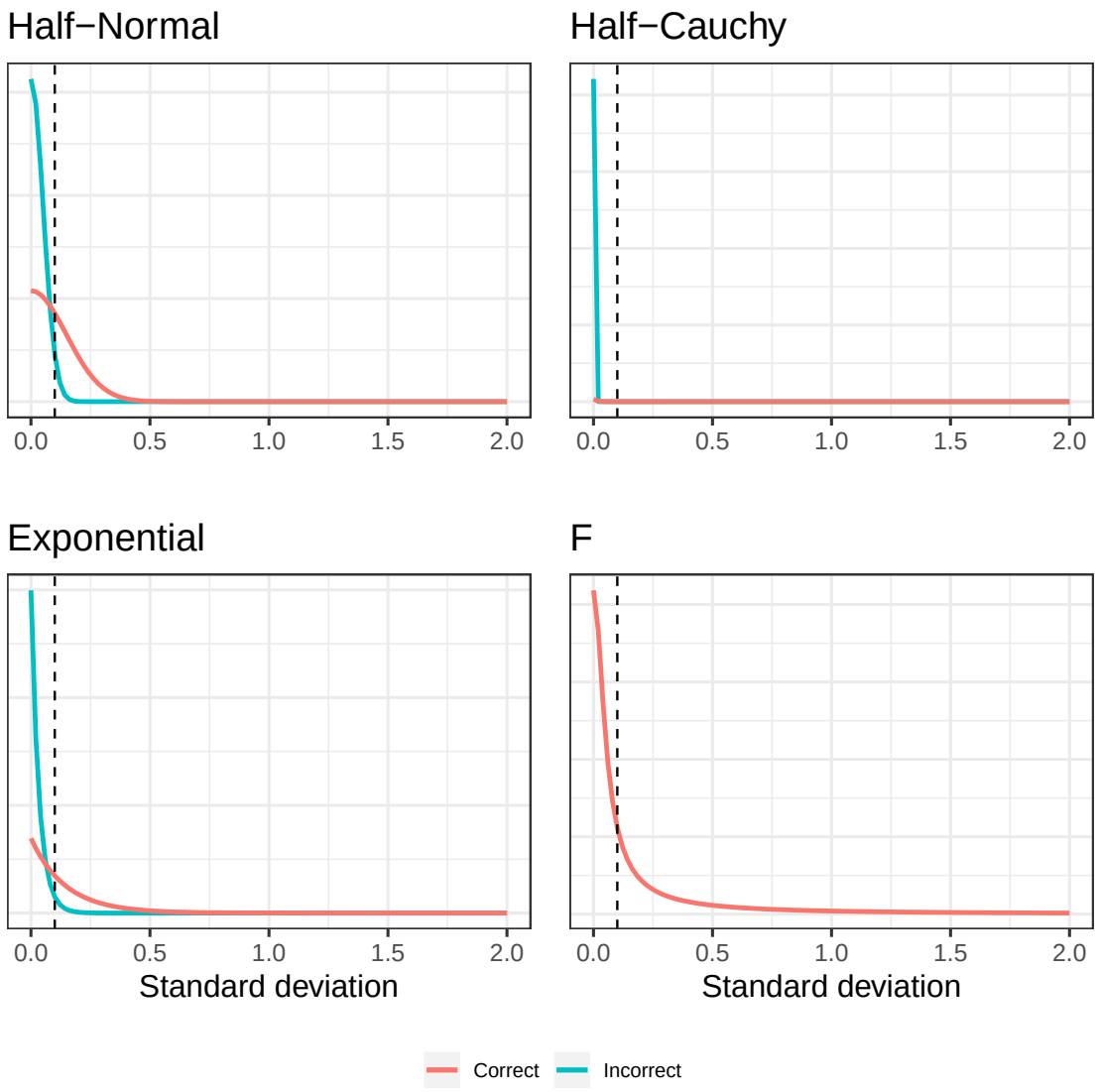

Figure 3.6: Prior densities for the informative specifications when the population value for the standard deviation equals 0.1

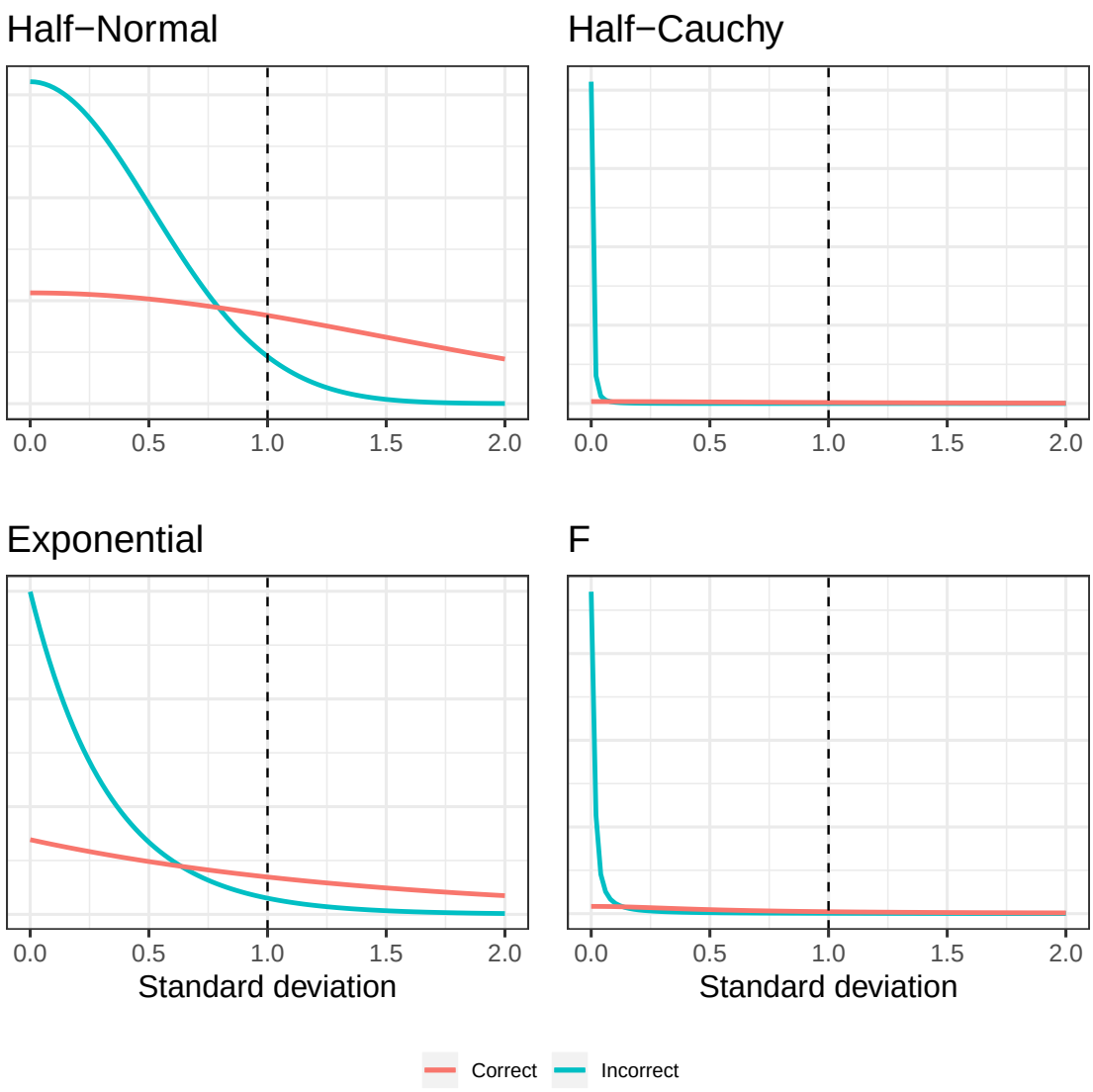

Figure 3.7: Prior densities for the informative specifications when the population value for the standard deviation equals 1

Study 3: Influence number of groups

Since the influence of the prior distribution decreases as the sample size increases, we vary the number of groups in the third study to assess when the results are no longer dependent on the prior distribution. Specifically, we consider a sample size of 50 groups with 20 observations within each group [4]. We do so for a subset of the population conditions from Study 1 to keep the simulation feasible. Specifically, we select those conditions with an effect size equal to -0.2, to be able to compute the power with an increasing number of groups, and a VPC equal to 0.03, since we

[4]We also ran the simulation with 100 groups and 20 observations within each group. However due to convergence issues and the fact that these results generally did not differ substantially from those of $G = 50$, we do not include them here. The results for $G = 100$ are available online at https://osf.io/pq8gm/.

expect the low VPC condition to be more problematic (see e.g., Hox & Maas, 2001). This leads to 4 different conditions, which are analysed using the 8 prior settings from Study 1, as well as the 4 correct informative priors from Study 2. Table 3.3 presents an overview of all the different priors that are considered in the various simulation study, as well as the abbreviations used to describe their results.

Table 3.3: Overview prior distributions investigated in the simulation studies.

Prior	Abbreviation	Included in study
Uniform	UN	1 & 3
inverse-Gamma(0.1, 0.1)	IG.1	1 & 3
inverse-Gamma(0.01, 0.01)	IG.01	1 & 3
inverse-Gamma(0.001, 0.001)	IG.001	1 & 3
default half-Normal	HNdef	1 & 3
default half-Cauchy	HCdef	1 & 3
default F	Fdef	1 & 3
default Exponential	EXPdef	1 & 3
correct informative half-Normal	HNinf	2 & 3
correct informative half-Cauchy	HCinf	2 & 3
correct informative F	Finf	2 & 3
correct informative Exponential	EXPinf	2 & 3
incorrect informative half-Normal	HNinc	2
incorrect informative half-Cauchy	HCinc	2
incorrect informative F	Finc	2
incorrect informative Exponential	EXPinc	2

Outcomes

The specific point estimate used to summarize the posterior distribution can influence the results (see e.g., Browne & Draper, 2006). Therefore, we consider the posterior mean, median, and mode when computing the various outcomes. The outcomes we consider are [5]:

1. ***Bias.*** We consider the bias, computed as $\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i - \theta$, where N denotes the number of replications, θ is the population value for the parameter of

[5]Initially, we also compared the precision of the various priors through the empirical SE, which is computed as $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_i - \bar{\theta})^2}$, with $\bar{\theta}$ denoting the mean of $\hat{\theta}_i$. However, the empirical SE was sufficiently small in all conditions and for all priors (in the order of 1e-14) that we do not include this outcome in the results.

interest used to generate the simulation data, and $\hat{\theta}_i$ is the posterior summary for that parameter in replication i. When $|\theta| > 0$, we will also consider the relative percent bias, which facilitates more straightforward comparisons across population values.

2. ***Mean squared error (MSE).*** As a composite measure of the bias and variance, we consider the MSE. The MSE is computed as: $\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2$.

3. ***Coverage.*** Coverage rates are investigated by computing how often the population value is contained in the 95% credible interval.

4. ***Power/type 1 error.*** For the parameter of interest, the contextual effect, we investigate the power and type 1 error. The power and type 1 error are computed as the proportion of replications for which zero is not included in the 95% credible interval when the population ES equals -0.2 and 0, respectively.

We use 500 replications per condition and have computed the Monte Carlo SE for every outcome measure to quantify the uncertainty in the simulation results (Morris, White, & Crowther, 2019). All Monte Carlo SEs were sufficiently small to conclude that 500 replications was enough (the maximum MCSE was 0.03 for the coverage percentages).

All analyses are run using the R interface to Stan, `RStan` (Stan Development Team, 2018). For each analysis, we ran two MCMC chains with 3000 iterations each, half of which was used as burnin[6]. We used a maximum treedepth of 10 and set `adapt_delta` to 0.90. We compared various convergence criteria. The results presented below are based on "strict" convergence criteria in which a replication is considered converged if the split $\hat{R}$ (a version of the potential scale reduction factor, PSRF; Gelman and Rubin (1992)) is smaller than or equal to 1.05, there are no divergent transitions, and the maximum treedepth is not exceeded. We removed all replications that did not meet these criteria and present the results for those conditions with at least 50% convergence. However, we also checked whether the results differ if they are based only on those replications that are converged for every prior distribution to ensure comparability across priors. Unfortunately, given the strict convergence criteria, there was not much overlap in converged replications across priors and as a result most conditions included less than 100 converged replications. Therefore, we also consider "weak" convergence criteria in which a replication is considered converged if the maximum Rhat is smaller than or equal to 1.20, there is a maximum of 10 divergent transitions, and the maximum treedepth is not exceeded. These criteria generally led to sufficient overlap in converged replications

[6]To assess the influence of the number of iterations on convergence, we reran one condition with 6000 iterations. However, this did not increase the convergence percentage.

across priors. Here, we present only a selection of the results. Specifically, we report the results based on all converged replications according to the strict criteria and we will note if these results differ from those obtained using only the replications converged across all priors based on the weak criteria. We do not report the results for the power and type 1 error rates. Across all conditions, the type 1 error rates were generally close to the nominal 5% whereas the power to find a small contextual effect was much too low, ranging from 0.01 to 0.09 even for 50 groups. Additional results from the simulation study, including the type 1 error rates and power, as well as all the code can be found online at https://osf.io/pq8gm/. Please note that throughout the tables and figures below, we use the abbreviations for the priors presented in Table 3.3.

Results Study 1: Influence population values

Convergence

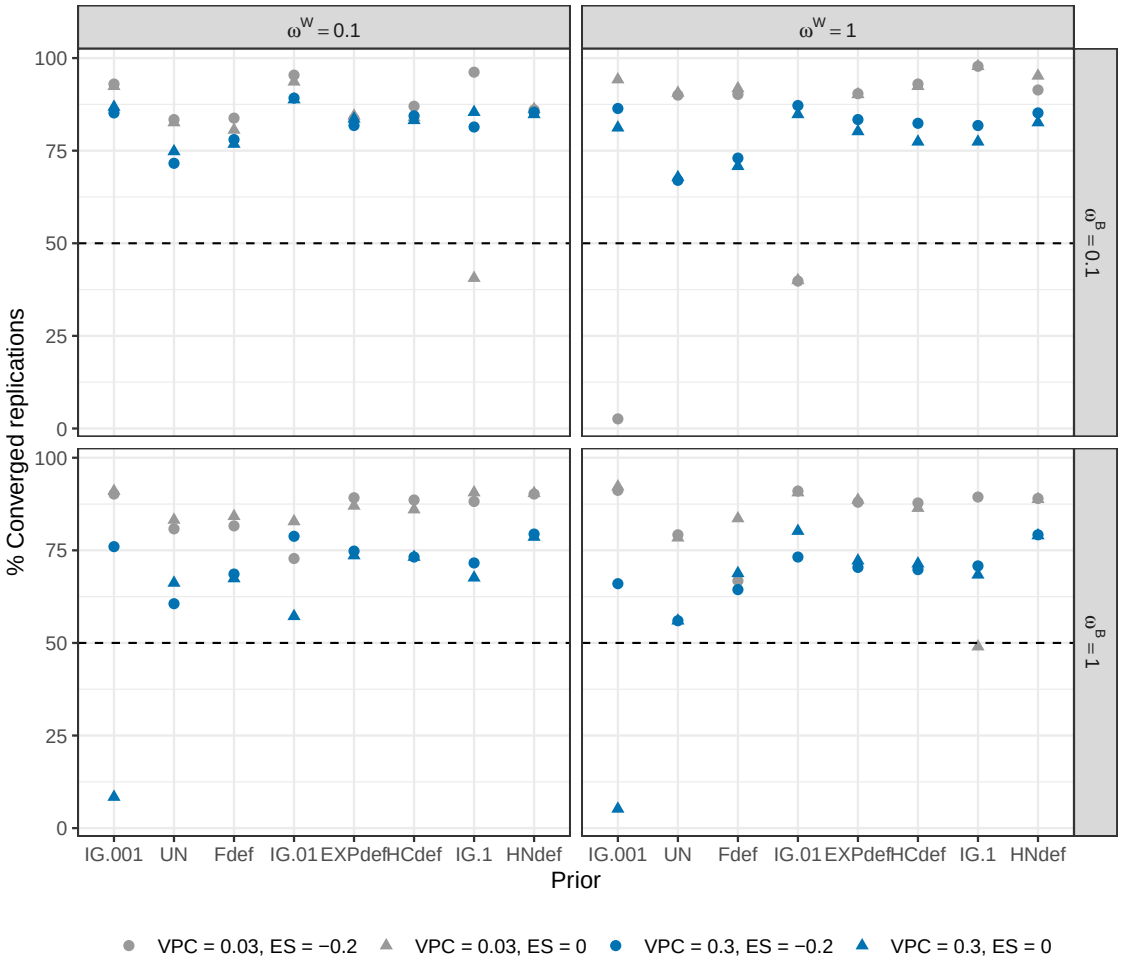

Figure 3.8: Percentages converged replications per condition according to the strict convergence criteria

Figure 3.8 shows the percentages of converged replications according to the strict criteria for each prior in the first study. The x-axis is ordered such that the prior on the left has the lowest convergence percentage across conditions, while the prior on the right has the highest convergence percentage. The dashed lines indicate 50% convergence and conditions falling below this line are not included in the results. This is the case in seven conditions, specifically for the inverse-Gamma priors. Generally, there are only slight differences in convergence for varying effect sizes (circles vs. triangles). With regard to the VPC, the smaller VPC of 0.03 generally leads to higher convergence compared to the larger VPC of 0.3 (grey vs. blue).

Bias

Table 3.4 shows the relative bias with the absolute bias in brackets for the variance parameters and the parameter of interest, $\beta_B - \beta_W$, based on the posterior median estimates and strict convergence criteria. In general, if the posterior estimates differ, the modes showed the least bias, followed by the posterior median and finally the posterior mean. However, the posterior mode is computed in a slightly adhoc manner using kernel density estimation. Therefore, we instead base the outcomes on the posterior median estimates. Furthermore, since the results do not differ substantially across effect sizes, we report the bias only for those conditions where $ES = -0.2$. The full results are available online.

For the standard deviation of the latent variable at the between level, ω_B, all priors result in some bias, although the values are comparable across most priors with the exception of the inverse-Gamma priors. Most priors underestimate ω_B when the population value equals 1 and when the population value equals 0.1 combined with a VPC of 0.03. The IG(0.1, 0.1) prior, however, largely overestimates ω_B when the population value equals 0.1, regardless of the value of the VPC.

For the standard deviation of the latent variable at the within level, ω_W, the robust priors show only small biases when the population values equal 0.1, with more substantial underestimation when the population values equal 1, whereas the IG(0.1, 0.1) and IG(0.01, 0.01) priors show a substantial overestimation when the population values equal 0.01.

For the standard deviation of the items at the between level, σ_Y^B, we see a similar picture across all items. Specifically, the bias is generally small except for the IG(0.1, 0.1) prior when $\omega_B = 0.1$ and $VPC = 0.03$.

All priors underestimate the parameter of interest, the contextual effect $\beta_B - \beta_W$, regardless of the population condition. The relative bias is close to negative 100% when ω_W equals 1 for all priors, but differs more across the priors when ω_W equals 0.1.

Table 3.4: Relative bias with absolute bias in brackets for selected parameters based on strict convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
						Standard deviation latent variable between ω_B					
0.03	0.1	0.1	-0.2	-0.427 (-0.043)	-0.393 (-0.039)	0.189 (0.019)	1.624 (0.162)	-0.429 (-0.043)	-0.435 (-0.043)	-0.428 (-0.043)	-0.46 (-0.046)
0.03	0.1	1	-0.2	-0.342 (-0.034)	NA	NA	1.577 (0.158)	-0.351 (-0.035)	-0.356 (-0.036)	-0.358 (-0.036)	-0.39 (-0.039)
0.03	1	0.1	-0.2	-0.834 (-0.834)	-0.876 (-0.876)	-0.801 (-0.801)	-0.609 (-0.609)	-0.827 (-0.827)	-0.832 (-0.832)	-0.829 (-0.829)	-0.842 (-0.842)
0.03	1	1	-0.2	-0.84 (-0.84)	-0.882 (-0.882)	-0.814 (-0.814)	-0.637 (-0.637)	-0.835 (-0.835)	-0.84 (-0.84)	-0.838 (-0.838)	-0.85 (-0.85)
0.3	0.1	0.1	-0.2	0.662 (0.066)	0.245 (0.024)	1.029 (0.103)	2.962 (0.296)	0.746 (0.075)	0.693 (0.069)	0.703 (0.07)	0.561 (0.056)
0.3	0.1	1	-0.2	0.948 (0.095)	0.32 (0.032)	1.094 (0.109)	2.985 (0.298)	1.046 (0.105)	0.906 (0.091)	1 (0.1)	0.83 (0.083)
0.3	1	0.1	-0.2	-0.772 (-0.772)	-0.843 (-0.843)	-0.758 (-0.758)	-0.546 (-0.546)	-0.755 (-0.755)	-0.773 (-0.773)	-0.771 (-0.771)	-0.785 (-0.785)
0.3	1	1	-0.2	-0.759 (-0.759)	-0.837 (-0.837)	-0.758 (-0.758)	-0.555 (-0.555)	-0.734 (-0.734)	-0.749 (-0.749)	-0.749 (-0.749)	-0.768 (-0.768)
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
						Standard deviation latent variable within ω_W					
0.03	0.1	0.1	-0.2	0.119 (0.012)	-0.179 (-0.018)	0.669 (0.067)	2.671 (0.267)	0.096 (0.01)	0.053 (0.005)	0.097 (0.01)	-0.017 (-0.002)
0.03	0.1	1	-0.2	-0.818 (-0.818)	NA	NA	-0.584 (-0.584)	-0.826 (-0.826)	-0.833 (-0.833)	-0.827 (-0.827)	-0.842 (-0.842)
0.03	1	0.1	-0.2	0.258 (0.026)	-0.197 (-0.02)	0.507 (0.051)	2.145 (0.215)	0.129 (0.013)	0.089 (0.009)	0.133 (0.013)	-0.001 (0)
0.03	1	1	-0.2	-0.814 (-0.814)	-0.89 (-0.89)	-0.821 (-0.821)	-0.651 (-0.651)	-0.835 (-0.835)	-0.842 (-0.842)	-0.837 (-0.837)	-0.853 (-0.853)
0.3	0.1	0.1	-0.2	0.227 (0.023)	-0.2 (-0.02)	0.481 (0.048)	2.085 (0.208)	0.072 (0.007)	0.05 (0.005)	0.094 (0.009)	-0.023 (-0.002)
0.3	0.1	1	-0.2	-0.819 (-0.819)	-0.892 (-0.892)	-0.825 (-0.825)	-0.667 (-0.667)	-0.849 (-0.849)	-0.845 (-0.845)	-0.837 (-0.837)	-0.859 (-0.859)
0.3	1	0.1	-0.2	0.297 (0.03)	-0.158 (-0.016)	0.481 (0.048)	2.032 (0.203)	0.084 (0.008)	0.103 (0.01)	0.174 (0.017)	0.003 (0)
0.3	1	1	-0.2	-0.814 (-0.814)	-0.89 (-0.89)	-0.822 (-0.822)	-0.669 (-0.669)	-0.848 (-0.848)	-0.843 (-0.843)	-0.832 (-0.832)	-0.859 (-0.859)
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
						Standard deviation item 1 between σ^B_{y1}					
0.03	0.1	0.1	-0.2	-0.008 (-0.003)	-0.248 (-0.083)	-0.092 (-0.031)	0.187 (0.062)	-0.036 (-0.012)	-0.051 (-0.017)	-0.032 (-0.011)	-0.068 (-0.023)
0.03	0.1	1	-0.2	-0.038 (-0.014)	NA	NA	0.112 (0.042)	-0.077 (-0.029)	-0.091 (-0.035)	-0.082 (-0.031)	-0.122 (-0.046)
0.03	1	0.1	-0.2	0.04 (0.042)	0 (0)	0.005 (0.005)	-0.016 (-0.017)	-0.023 (-0.024)	-0.004 (-0.005)	-0.001 (-0.001)	-0.013 (-0.014)
0.03	1	1	-0.2	0.02 (0.022)	-0.037 (-0.039)	-0.035 (-0.037)	-0.031 (-0.033)	-0.046 (-0.048)	-0.039 (-0.042)	-0.035 (-0.037)	-0.034 (-0.036)
0.3	0.1	0.1	-0.2	0.033 (0.04)	-0.015 (-0.017)	-0.01 (-0.012)	-0.014 (-0.017)	-0.034 (-0.04)	-0.025 (-0.029)	-0.003 (-0.003)	-0.018 (-0.022)
0.3	0.1	1	-0.2	0.046 (0.063)	-0.005 (-0.007)	-0.003 (-0.004)	-0.002 (-0.003)	-0.047 (-0.065)	-0.019 (-0.026)	-0.008 (-0.01)	-0.027 (-0.037)
0.3	1	0.1	-0.2	0.054 (0.083)	0.01 (0.015)	0.019 (0.03)	0.022 (0.035)	-0.049 (-0.077)	-0.012 (-0.019)	0.019 (0.03)	-0.014 (-0.022)
0.3	1	1	-0.2	0.021 (0.036)	-0.009 (-0.014)	-0.022 (-0.037)	-0.016 (-0.028)	-0.093 (-0.157)	-0.037 (-0.062)	-0.017 (-0.029)	-0.049 (-0.083)
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
						Contextual effect β_B - β_W					
0.03	0.1	0.1	-0.2	-1.132 (0.016)	-1.141 (0.016)	-1.266 (0.018)	-1.435 (0.02)	-1.103 (0.015)	-1.115 (0.016)	-1.165 (0.016)	-1.122 (0.016)
0.03	0.1	1	-0.2	-0.982 (0.102)	NA	NA	-0.957 (0.1)	-0.969 (0.101)	-0.99 (0.103)	-0.985 (0.102)	-0.978 (0.102)
0.03	1	0.1	-0.2	-1.165 (0.016)	-1.118 (0.016)	-1.393 (0.019)	-1.078 (0.015)	-1.202 (0.017)	-1.179 (0.017)	-1.252 (0.018)	-1.238 (0.017)
0.03	1	1	-0.2	-0.979 (0.102)	-0.987 (0.103)	-0.949 (0.099)	-0.917 (0.095)	-0.99 (0.103)	-0.974 (0.101)	-0.933 (0.097)	-0.981 (0.102)
0.3	0.1	0.1	-0.2	-0.693 (0.01)	-0.745 (0.01)	-0.718 (0.01)	-0.479 (0.007)	-0.831 (0.012)	-0.601 (0.008)	-0.878 (0.012)	-0.742 (0.01)
0.3	0.1	1	-0.2	-0.976 (0.102)	-1.013 (0.105)	-0.997 (0.104)	-1.083 (0.113)	-1.03 (0.107)	-1.019 (0.106)	-1.022 (0.106)	-1.017 (0.106)
0.3	1	0.1	-0.2	-1.01 (0.014)	-1.125 (0.016)	-1.248 (0.017)	-1.63 (0.023)	-1.219 (0.017)	-1.462 (0.02)	-1.172 (0.016)	-1.376 (0.019)
0.3	1	1	-0.2	-1.054 (0.11)	-1.037 (0.108)	-0.973 (0.101)	-1.066 (0.111)	-0.96 (0.1)	-1.007 (0.105)	-0.985 (0.102)	-1.03 (0.107)

Note. NA indicates that results are not available since the convergence percentage < 50% in this condition.

Mean squared error (MSE)

The mean squared error combines the influence of the bias and variance in one estimate. Note, however, that since the methods are not all unbiased, the relative influence of the bias and variance on the MSE can vary with the sample size (Morris et al., 2019).

Table 3.5 shows the MSE for the variance parameters and the parameter of interest, $\beta_B - \beta_W$, based on the posterior median estimates. Although the differences in MSE across posterior estimates were generally small, the median and mode showed the lowest MSE when differences arose. Again, we only report the results for the effect size of -0.2, since the results did not differ substantially across effect sizes. For ω_B, the MSE is close to zero for all priors when the population value equals

0.1, but slightly larger when the population value equals 1. In this case, the IG(0.1, 0.1) prior shows the smallest MSE, while the MSE is comparable across the other priors. A very similar picture arises for ω_W, with small MSE when the population value for ω_W equals 0.1 and larger MSE when the population value equals 1. For σ^B_{y1}, all priors result in comparable small MSEs, with slightly larger values when the VPC equals 0.3 combined with one or both of the population values for ω equal to 1. Finally, for the contextual effect $\beta_B - \beta_W$ the MSE is close to zero across priors and population values.

Table 3.5: Mean squared error (MSE) for selected parameters based on strict convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable between ω_B											
0.03	0.1	0.1	-0.2	0.002	0.002	0.001	0.027	0.003	0.003	0.002	0.003
0.03	0.1	1	-0.2	0.002	NA	NA	0.026	0.002	0.002	0.002	0.002
0.03	1	0.1	-0.2	0.701	0.769	0.644	0.378	0.690	0.698	0.693	0.714
0.03	1	1	-0.2	0.711	0.779	0.664	0.410	0.701	0.709	0.707	0.726
0.3	0.1	0.1	-0.2	0.011	0.003	0.014	0.094	0.012	0.010	0.011	0.007
0.3	0.1	1	-0.2	0.018	0.004	0.015	0.096	0.019	0.014	0.018	0.014
0.3	1	0.1	-0.2	0.603	0.715	0.579	0.307	0.579	0.605	0.601	0.622
0.3	1	1	-0.2	0.588	0.708	0.581	0.318	0.551	0.575	0.574	0.603
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable within ω_W											
0.03	0.1	0.1	-0.2	0.004	0.001	0.006	0.076	0.003	0.003	0.003	0.002
0.03	0.1	1	-0.2	0.677	NA	NA	0.349	0.690	0.700	0.692	0.715
0.03	1	0.1	-0.2	0.005	0.001	0.004	0.049	0.004	0.003	0.004	0.003
0.03	1	1	-0.2	0.673	0.795	0.677	0.429	0.705	0.715	0.707	0.733
0.3	0.1	0.1	-0.2	0.004	0.001	0.003	0.046	0.003	0.002	0.003	0.002
0.3	0.1	1	-0.2	0.681	0.797	0.684	0.449	0.726	0.720	0.708	0.742
0.3	1	0.1	-0.2	0.004	0.001	0.003	0.043	0.003	0.002	0.003	0.002
0.3	1	1	-0.2	0.674	0.795	0.677	0.452	0.725	0.716	0.700	0.742
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation item 1 between σ^B_{y1}											
0.03	0.1	0.1	-0.2	0.021	0.029	0.015	0.011	0.020	0.020	0.020	0.021
0.03	0.1	1	-0.2	0.023	NA	NA	0.011	0.023	0.022	0.022	0.024
0.03	1	0.1	-0.2	0.074	0.067	0.077	0.074	0.053	0.063	0.066	0.060
0.03	1	1	-0.2	0.063	0.060	0.061	0.061	0.051	0.052	0.057	0.052
0.3	0.1	0.1	-0.2	0.074	0.073	0.076	0.085	0.057	0.067	0.074	0.066
0.3	0.1	1	-0.2	0.109	0.096	0.097	0.105	0.070	0.088	0.088	0.081
0.3	1	0.1	-0.2	0.140	0.126	0.126	0.132	0.077	0.106	0.119	0.104
0.3	1	1	-0.2	0.141	0.128	0.136	0.134	0.106	0.130	0.132	0.114
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Contextual effect β_B - β_W											
0.03	0.1	0.1	-0.2	0.002	0.002	0.004	0.012	0.002	0.002	0.002	0.002
0.03	0.1	1	-0.2	0.013	NA	NA	0.020	0.013	0.013	0.013	0.013
0.03	1	0.1	-0.2	0.008	0.005	0.008	0.024	0.007	0.007	0.007	0.007
0.03	1	1	-0.2	0.016	0.014	0.016	0.027	0.016	0.016	0.014	0.016
0.3	0.1	0.1	-0.2	0.007	0.004	0.007	0.021	0.006	0.006	0.006	0.005
0.3	0.1	1	-0.2	0.020	0.017	0.019	0.038	0.020	0.020	0.021	0.019
0.3	1	0.1	-0.2	0.012	0.008	0.012	0.033	0.011	0.010	0.011	0.009
0.3	1	1	-0.2	0.025	0.019	0.030	0.037	0.021	0.021	0.022	0.021

Note. NA indicates that results are not available since the convergence percentage < 50% in this condition.

Coverage

Table 3.6 shows the coverage rates based on the 95% credibility intervals. For ω_B and ω_W, the coverage rates are above the nominal 95% when the corresponding population values equal 0.1 for all priors except the IG(0.1, 0.1), which results in coverage rates of 0. When the population values equal 1, all coverage rates are much too low. For σ^B_{y1}, all coverage rates are close to the nominal 95% and are therefore not reported. Finally, for the parameter of interest, the coverage rates are close to 95% when the effect size equals 0. When the effect size equals -0.2, the coverage rates are close to 95% only when $\omega_W = 0.1$ and they are generally too low when $\omega_W = 1$, except for the IG(0.1, 0.1) prior.

Table 3.6: 95% coverage for selected parameters based on strict convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable between ω_B											
0.03	0.1	0.1	-0.2	100.0	100.0	100.0	0.0	100.0	100.0	100.0	100.0
0.03	0.1	0.1	0	100.0	100.0	99.8	NA	100.0	100.0	100.0	100.0
0.03	0.1	1	-0.2	100.0	NA	NA	0.0	100.0	100.0	100.0	100.0
0.03	0.1	1	0	100.0	100.0	NA	0.0	100.0	100.0	100.0	100.0
0.03	1	0.1	-0.2	10.9	1.1	5.2	24.5	6.4	6.8	5.9	3.8
0.03	1	0.1	0	10.8	1.3	5.1	27.4	6.2	6.0	6.4	5.5
0.03	1	1	-0.2	9.6	0.2	1.8	14.8	4.5	3.6	4.2	2.7
0.03	1	1	0	9.4	1.5	3.1	NA	4.7	4.9	5.7	4.1
0.3	0.1	0.1	-0.2	98.9	99.1	93.0	0.0	97.7	98.1	97.9	98.5
0.3	0.1	0.1	0	99.7	99.8	92.1	0.0	99.3	99.8	99.5	99.5
0.3	0.1	1	-0.2	96.7	97.9	90.6	0.0	96.2	97.1	95.9	96.6
0.3	0.1	1	0	99.1	99.3	94.1	0.0	97.1	98.2	98.3	98.0
0.3	1	0.1	-0.2	32.3	5.0	11.2	59.5	16.1	14.8	17.5	12.6
0.3	1	0.1	0	33.2	NA	15.4	62.1	22.1	16.9	20.5	12.0
0.3	1	1	-0.2	37.5	3.9	11.2	53.1	24.5	21.8	23.6	17.0
0.3	1	1	0	42.5	NA	14.7	62.3	24.8	21.6	29.9	17.7
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable within ω_W											
0.03	0.1	0.1	-0.2	99.5	99.8	97.9	0.0	99.5	99.5	99.5	99.5
0.03	0.1	0.1	0	99.3	100.0	98.1	NA	99.5	99.5	99.3	99.3
0.03	0.1	1	-0.2	10.2	NA	NA	27.2	3.3	3.0	3.5	2.4
0.03	0.1	1	0	6.6	0.4	NA	25.8	2.3	2.4	3.3	1.6
0.03	1	0.1	-0.2	99.0	99.8	99.2	0.0	99.3	99.8	99.8	100.0
0.03	1	0.1	0	99.8	100.0	99.5	0.0	99.8	99.8	99.8	100.0
0.03	1	1	-0.2	8.8	0.9	3.1	8.7	4.3	3.4	4.5	3.6
0.03	1	1	0	6.1	0.2	0.9	NA	1.8	1.6	2.2	1.1
0.3	0.1	0.1	-0.2	99.7	100.0	99.8	0.0	99.8	99.8	99.7	99.8
0.3	0.1	0.1	0	99.7	100.0	99.3	0.0	100.0	99.8	99.5	100.0
0.3	0.1	1	-0.2	6.9	0.2	1.4	6.4	2.1	1.9	3.8	1.7
0.3	0.1	1	0	5.3	0.0	0.5	5.4	0.2	1.0	2.3	0.2
0.3	1	0.1	-0.2	99.7	100.0	99.2	0.0	99.7	99.7	99.4	100.0
0.3	1	0.1	0	100.0	NA	100.0	0.0	100.0	100.0	100.0	100.0
0.3	1	1	-0.2	11.1	0.6	1.1	5.6	2.3	2.0	4.7	1.4
0.3	1	1	0	7.9	NA	1.0	5.6	1.3	1.4	2.0	1.1
VPC	ω_B	ω_W	ES	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Contextual effect β_B - β_W											
0.03	0.1	0.1	-0.2	96.6	95.9	97.1	98.3	96.7	97.2	96.2	96.4
0.03	0.1	0.1	0	97.3	97.0	97.9	NA	97.0	97.9	98.3	97.2
0.03	0.1	1	-0.2	67.8	NA	NA	92.8	65.4	62.2	65.0	62.2
0.03	0.1	1	0	95.1	95.1	NA	98.0	95.6	95.7	95.6	95.6
0.03	1	0.1	-0.2	95.0	93.8	94.5	96.1	94.5	94.4	95.1	94.4
0.03	1	0.1	0	91.3	91.6	92.8	93.6	92.5	91.2	92.2	91.0
0.03	1	1	-0.2	87.9	74.6	84.2	94.2	85.4	84.7	87.7	82.7
0.03	1	1	0	93.9	93.7	94.0	NA	94.8	93.8	94.7	94.6
0.3	0.1	0.1	-0.2	96.1	95.1	96.9	97.1	96.0	96.2	96.2	95.8
0.3	0.1	0.1	0	97.9	96.1	96.8	97.4	95.8	96.9	97.4	97.4
0.3	0.1	1	-0.2	88.4	79.4	86.7	91.9	86.2	86.4	87.1	87.3
0.3	0.1	1	0	97.6	96.3	97.6	97.2	98.1	97.2	97.7	97.8
0.3	1	0.1	-0.2	96.0	95.3	95.9	95.8	94.7	94.8	96.2	95.5
0.3	1	0.1	0	96.1	NA	95.5	96.4	96.2	94.8	96.1	95.1
0.3	1	1	-0.2	89.3	81.8	88.0	94.6	88.9	87.4	89.8	88.1
0.3	1	1	0	95.0	NA	95.3	96.2	94.2	95.8	95.6	96.1

Note. NA indicates that results are not available since the convergence percentage < 50% in this condition.

Results Study 2: Influence prior misspecifications

We now turn to the results of the second simulation study in which we investigate the same population conditions as in Study 1, but with informative priors. The informative priors are specified such that they are either in line with the population values, or not.

Convergence

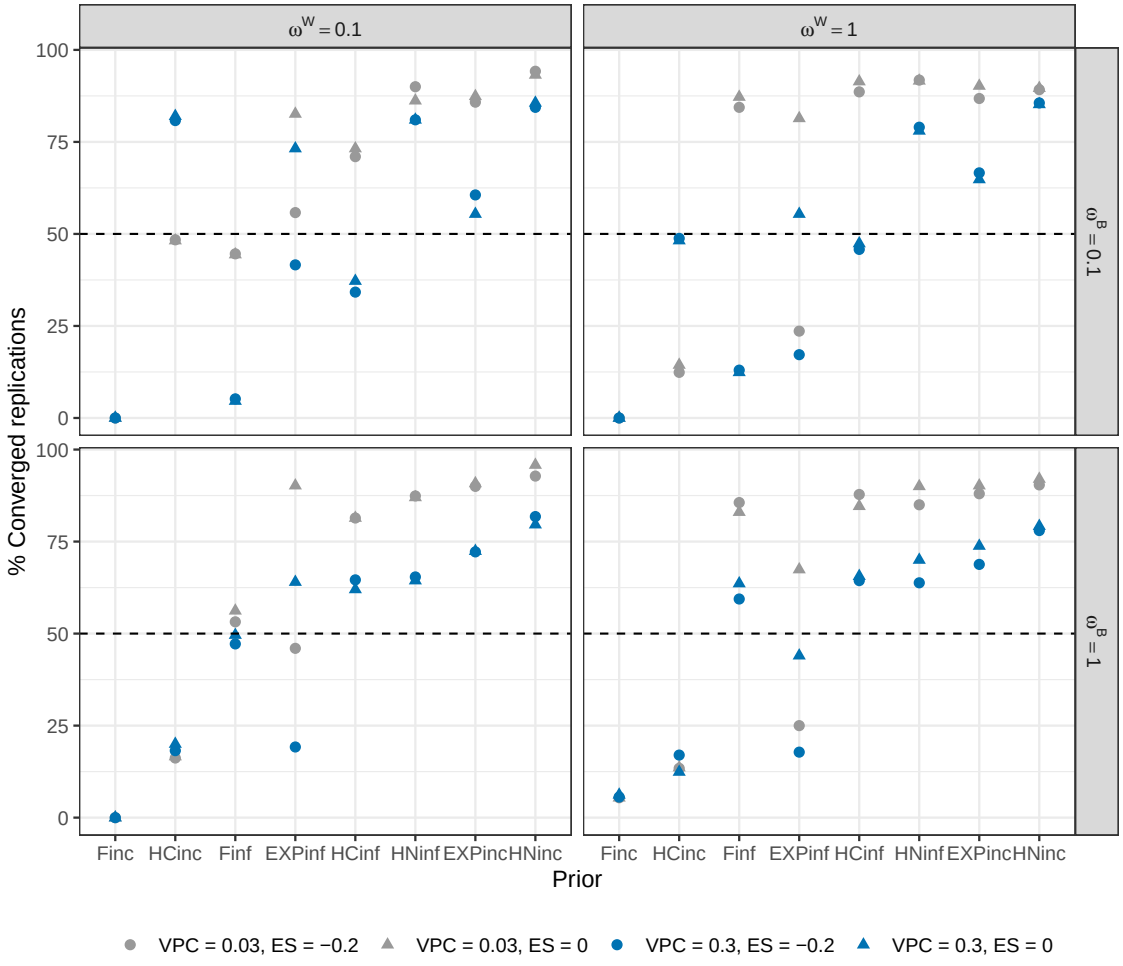

Figure 3.9: Percentages converged replications per condition according to the strict convergence criteria

Figure 3.9 shows the percentages of converged replications according to the strict criteria for each prior in the second study. The x-axis is again ordered such that the prior on the left has the lowest convergence percentage across conditions and the dashed lines indicate 50% convergence with conditions falling below this line are not included in the results. It is clear that, compared to the default priors in Study 1, the informative priors have more convergence problems. This is especially the case for the incorrectly specified F and half-Cauchy priors. For the incorrectly

specified F prior, the convergence of 0% when at least one of the population standard deviations equals 0.1 is due to the fact that in these situations the prior scale equals zero whereas it should be positive (see also Subsection 3.6). Only the informative half-Normal priors and the incorrectly specified Exponential prior obtained more than 50% convergence in all conditions.

In order to be able to compare the informative priors, we base the results of the second simulation study not on the strict convergence criteria, but on the weak convergence criteria. The convergence percentages based on these criteria are shown in Figure 3.10 and are much better compared to the strict convergence criteria. Although the average results do not differ substantially between the strict and weak convergence criteria, this choice does complicate comparisons between the first and second simulation study. The full results according to the strict criteria are available online at https://osf.io/pq8gm/.

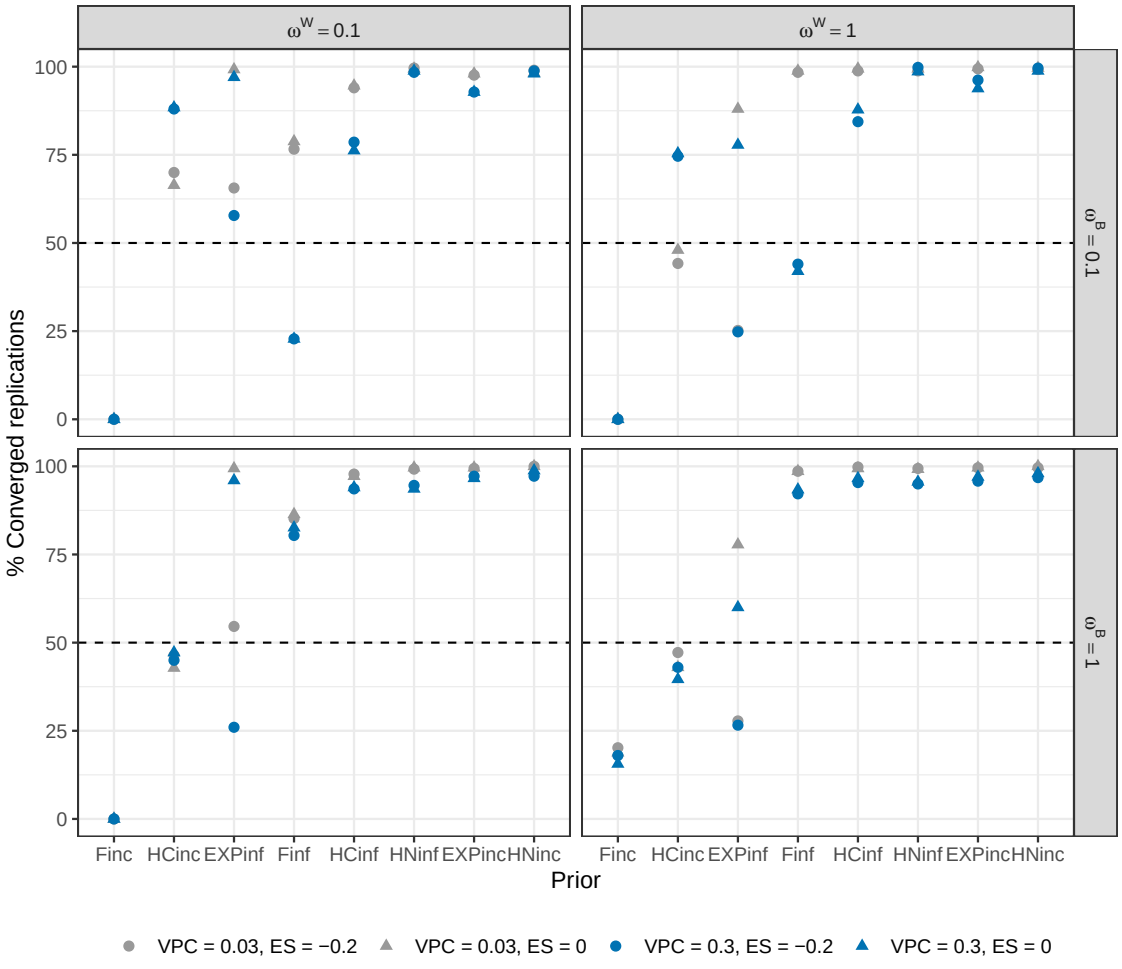

Figure 3.10: Percentages converged replications per condition according to the weak convergence criteria

Bias

Table 3.7 shows the relative bias with the absolute bias in brackets for the variance parameters and the parameter of interest, $\beta_B - \beta_W$, based on the posterior median estimates and weak convergence criteria. Again, we only report the bias for those conditions in which $ES = -0.2$ and refer to the online materials for the full results.

As expected, the bias is generally lower for the correctly specified informative priors compared to the incorrectly specified informative priors. However, the value of the difference between the priors varies across conditions. For example, for ω_B, the informative half-normal prior only shows a substantially lower bias when the $VPC = 0.3$ and $\omega_B = 0.1$. In these two conditions, the informative half-normal priors also outperform the default half-normal prior from Study 1. In the other conditions, the correctly specified informative half-normal prior performs comparably to the default half-normal prior. In general, it depends on the population condition whether the informative priors outperform the default priors in terms of bias.In theory, any differences in bias between the default and informative priors might be due to the difference in convergence criteria used. To check this, we compared the bias across default and informative priors using the weak criteria for both types and found that the differences in bias were due to the type of prior rather than the convergence criteria. For ω_W, we see that the smaller biases when the population values equal 0.1 that we found in Study 1 do not hold for all informative priors. We see this pattern for the informative half-normal prior and, to a lesser extent, the incorrect half-normal and exponential priors, but not for the informative half-Cauchy or F priors. These priors show substantial biases regardless of the population values. The biases for σ^B_{y1} are substantial especially when the $VPC = 0.03$ and $\omega_B = 0.1$, while the contextual effect $\beta_B - \beta_W$ is again underestimated across the board.

Table 3.7: Relative bias with absolute bias in brackets for selected parameters based on weak convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
						Standard deviation latent variable between ω_B				
0.03	0.1	0.1	-0.2	-0.563 (-0.056)	-0.651 (-0.065)	-0.873 (-0.087)	-0.992 (-0.099)	-0.916 (-0.092)	-0.617 (-0.062)	-0.721 (-0.072)
0.03	0.1	1	-0.2	-0.471 (-0.047)	-0.61 (-0.061)	-0.873 (-0.087)	NA	-0.906 (-0.091)	NA	-0.689 (-0.069)
0.03	1	0.1	-0.2	-0.859 (-0.859)	-0.851 (-0.851)	-0.873 (-0.873)	NA	-0.878 (-0.878)	-0.865 (-0.865)	-0.868 (-0.868)
0.03	1	1	-0.2	-0.841 (-0.841)	-0.827 (-0.827)	-0.843 (-0.843)	NA	-0.87 (-0.87)	NA	-0.856 (-0.856)
0.3	0.1	0.1	-0.2	-0.123 (-0.012)	-0.534 (-0.053)	-0.887 (-0.089)	-0.996 (-0.1)	NA	-0.159 (-0.016)	-0.625 (-0.063)
0.3	0.1	1	-0.2	-0.025 (-0.002)	-0.552 (-0.055)	-0.885 (-0.088)	-0.99 (-0.099)	NA	NA	-0.643 (-0.064)
0.3	1	0.1	-0.2	-0.812 (-0.812)	-0.809 (-0.809)	-0.837 (-0.837)	NA	-0.852 (-0.852)	NA	-0.834 (-0.834)
0.3	1	1	-0.2	-0.765 (-0.765)	-0.763 (-0.763)	-0.784 (-0.784)	NA	-0.834 (-0.834)	NA	-0.806 (-0.806)
VPC	ω_B	ω_W	ES	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
						Standard deviation latent variable within ω_W				
0.03	0.1	0.1	-0.2	-0.309 (-0.031)	-0.674 (-0.067)	-0.898 (-0.09)	-0.999 (-0.1)	-0.96 (-0.096)	-0.41 (-0.041)	-0.772 (-0.077)
0.03	0.1	1	-0.2	-0.846 (-0.846)	-0.89 (-0.89)	-0.88 (-0.88)	NA	-0.9 (-0.9)	NA	-0.906 (-0.906)
0.03	1	0.1	-0.2	-0.247 (-0.025)	-0.669 (-0.067)	-0.898 (-0.09)	NA	-0.957 (-0.096)	-0.322 (-0.032)	-0.773 (-0.077)
0.03	1	1	-0.2	-0.823 (-0.823)	-0.868 (-0.868)	-0.838 (-0.838)	NA	-0.887 (-0.887)	NA	-0.895 (-0.895)
0.3	0.1	0.1	-0.2	-0.305 (-0.031)	-0.689 (-0.069)	-0.906 (-0.091)	-0.999 (-0.1)	NA	-0.392 (-0.039)	-0.783 (-0.078)
0.3	0.1	1	-0.2	-0.861 (-0.861)	-0.904 (-0.904)	-0.872 (-0.872)	-0.993 (-0.993)	NA	NA	-0.912 (-0.912)
0.3	1	0.1	-0.2	-0.207 (-0.021)	-0.664 (-0.066)	-0.9 (-0.09)	NA	-0.958 (-0.096)	NA	-0.77 (-0.077)
0.3	1	1	-0.2	-0.815 (-0.815)	-0.87 (-0.87)	-0.83 (-0.83)	NA	-0.884 (-0.884)	NA	-0.896 (-0.896)
VPC	ω_B	ω_W	ES	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
						Standard deviation item 1 between σ^B_{y1}				
0.03	0.1	0.1	-0.2	-0.11 (-0.037)	-0.48 (-0.161)	-0.421 (-0.141)	-0.976 (-0.327)	-0.577 (-0.193)	-0.124 (-0.042)	-0.555 (-0.186)
0.03	0.1	1	-0.2	-0.12 (-0.046)	-0.443 (-0.168)	-0.375 (-0.142)	NA	-0.533 (-0.201)	NA	-0.512 (-0.193)
0.03	1	0.1	-0.2	0.018 (0.019)	-0.116 (-0.121)	-0.001 (-0.001)	NA	-0.032 (-0.034)	0.005 (0.005)	-0.089 (-0.093)
0.03	1	1	-0.2	-0.01 (-0.01)	-0.137 (-0.146)	-0.03 (-0.031)	NA	-0.052 (-0.055)	NA	-0.116 (-0.123)
0.3	0.1	0.1	-0.2	0.017 (0.021)	-0.107 (-0.128)	0.013 (0.016)	-0.049 (-0.058)	NA	0.022 (0.026)	-0.087 (-0.103)
0.3	0.1	1	-0.2	0.022 (0.03)	-0.104 (-0.142)	0.014 (0.019)	-0.044 (-0.06)	NA	NA	-0.079 (-0.107)
0.3	1	0.1	-0.2	0.032 (0.05)	-0.093 (-0.144)	0.028 (0.043)	NA	-0.006 (-0.01)	NA	-0.064 (-0.099)
0.3	1	1	-0.2	0.006 (0.011)	-0.117 (-0.197)	0.006 (0.011)	NA	-0.032 (-0.054)	NA	-0.093 (-0.157)
VPC	ω_B	ω_W	ES	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
						Contextual effect β_B - β_W				
0.03	0.1	0.1	-0.2	-1.128 (0.016)	-1.106 (0.015)	-1.121 (0.016)	-1.228 (0.017)	-1.139 (0.016)	-1.072 (0.015)	-1.083 (0.015)
0.03	0.1	1	-0.2	-0.986 (0.103)	-0.99 (0.103)	-0.985 (0.102)	NA	-0.986 (0.103)	NA	-0.992 (0.103)
0.03	1	0.1	-0.2	-1.194 (0.017)	-1.152 (0.016)	-1.161 (0.016)	NA	-1.092 (0.015)	-1.136 (0.016)	-1.155 (0.016)
0.03	1	1	-0.2	-0.988 (0.103)	-0.979 (0.102)	-0.986 (0.103)	NA	-0.981 (0.102)	NA	-0.984 (0.102)
0.3	0.1	0.1	-0.2	-0.784 (0.011)	-0.896 (0.013)	-0.902 (0.013)	-0.883 (0.012)	NA	-0.427 (0.006)	-0.915 (0.013)
0.3	0.1	1	-0.2	-1.029 (0.107)	-1.019 (0.106)	-1.006 (0.105)	-1.016 (0.106)	NA	NA	-1.025 (0.107)
0.3	1	0.1	-0.2	-1.179 (0.016)	-1.237 (0.017)	-1.183 (0.017)	NA	-1.286 (0.018)	NA	-1.25 (0.017)
0.3	1	1	-0.2	-1.015 (0.106)	-1.008 (0.105)	-1.014 (0.105)	NA	-1.031 (0.107)	NA	-1.019 (0.106)

Note. Finc is missing due to complete nonconvergence. NA indicates that results are not available since the convergence percentage < 50% in this condition.

Mean squared error (MSE)

Table 3.8 shows the MSE for the variance parameters and the parameter of interest, $\beta_B - \beta_W$, based on the posterior median estimates and weak convergence criteria. The results are very similar to those of Study 1. For ω_B and ω_W, the MSE is close to zero for all priors when the corresponding population value equals 0.1, but slightly larger when the corresponding population value equals 1. There are no substantial differences across priors. All MSEs are small for σ^B_{y1} and generally comparable across population values and the MSEs for $\beta_B - \beta_W$ are close to zero

across the board. The MSEs do not show substantial differences between priors.

Table 3.8: Mean squared error (MSE) for selected parameters based on weak convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation latent variable between ω_B										
0.03	0.1	0.1	-0.2	0.003	0.008	0.009	0.004	0.004	0.010	0.005
0.03	0.1	1	-0.2	0.003	0.008	0.008	NA	0.004	NA	0.005
0.03	1	0.1	-0.2	0.742	0.764	0.773	0.751	0.726	NA	0.756
0.03	1	1	-0.2	0.712	0.715	0.759	NA	0.688	NA	0.736
0.3	0.1	0.1	-0.2	0.001	0.008	NA	0.001	0.003	0.010	0.004
0.3	0.1	1	-0.2	0.001	0.008	NA	NA	0.003	0.010	0.005
0.3	1	0.1	-0.2	0.664	0.704	0.729	NA	0.658	NA	0.698
0.3	1	1	-0.2	0.597	0.624	0.700	NA	0.590	NA	0.656
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation latent variable within ω_W										
0.03	0.1	0.1	-0.2	0.002	0.008	0.009	0.002	0.005	0.010	0.006
0.03	0.1	1	-0.2	0.721	0.778	0.812	NA	0.795	NA	0.823
0.03	1	0.1	-0.2	0.001	0.008	0.009	0.002	0.005	NA	0.006
0.03	1	1	-0.2	0.687	0.709	0.790	NA	0.758	NA	0.804
0.3	0.1	0.1	-0.2	0.001	0.008	NA	0.002	0.005	0.010	0.006
0.3	0.1	1	-0.2	0.747	0.764	NA	NA	0.820	0.986	0.834
0.3	1	0.1	-0.2	0.001	0.008	0.009	NA	0.004	NA	0.006
0.3	1	1	-0.2	0.675	0.697	0.785	NA	0.761	NA	0.805
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation item 1 between σ^B_{y1}										
0.03	0.1	0.1	-0.2	0.018	0.039	0.064	0.022	0.032	0.111	0.045
0.03	0.1	1	-0.2	0.021	0.042	0.071	NA	0.036	NA	0.051
0.03	1	0.1	-0.2	0.061	0.061	0.065	0.067	0.049	NA	0.056
0.03	1	1	-0.2	0.054	0.054	0.059	NA	0.052	NA	0.057
0.3	0.1	0.1	-0.2	0.065	0.065	NA	0.064	0.052	0.076	0.059
0.3	0.1	1	-0.2	0.086	0.085	NA	NA	0.067	0.090	0.073
0.3	1	0.1	-0.2	0.114	0.115	0.112	NA	0.077	NA	0.090
0.3	1	1	-0.2	0.122	0.125	0.120	NA	0.104	NA	0.111
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Contextual effect β_B - β_W										
0.03	0.1	0.1	-0.2	0.002	0.001	0.001	0.001	0.001	0.001	0.001
0.03	0.1	1	-0.2	0.013	0.012	0.012	NA	0.012	NA	0.012
0.03	1	0.1	-0.2	0.005	0.004	0.003	0.005	0.004	NA	0.004
0.03	1	1	-0.2	0.017	0.016	0.014	NA	0.015	NA	0.014
0.3	0.1	0.1	-0.2	0.004	0.003	NA	0.003	0.003	0.003	0.003
0.3	0.1	1	-0.2	0.018	0.017	NA	NA	0.016	0.014	0.016
0.3	1	0.1	-0.2	0.008	0.006	0.006	NA	0.007	NA	0.006
0.3	1	1	-0.2	0.023	0.022	0.019	NA	0.021	NA	0.019

Note. Finc is missing due to complete nonconvergence. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Coverage

Table 3.9 shows the coverage rates based on the 95% credibility intervals. It is clear that the coverage rates differ greatly across conditions and priors. For ω_B, the informative half-Normal and exponential priors show coverage rates above the nominal 95% when the $\omega_B = 0.1$ and much too low rates when $\omega_B = 1$. For the incorrect half-Normal and exponential priors, the coverage rates are only above 95% when $VPC = 0.3$ and $\omega_B = 0.1$ and they are too low otherwise. For the other informative priors, the rates are too low across all population conditions. For ω_W, the informative half-Normal and exponential priors show coverage rates of 100% when $\omega_W = 0.1$ and much too low rates otherwise. For the other priors, the coverage of ω_W is too low across the board, but especially when $\omega_W = 1$. Note that the coverage for the incorrectly specified half-Cauchy prior equals 0% for all conditions in which enough convergence was obtained to include the results. For σ^B_{y1}, the coverage rates are generally close to 95%, although they are slightly lower for some of the priors when $VPC = 0.03$ and $\omega_B = 0.1$. Again, the incorrect half-Cauchy prior shows very low coverage rates, but only when $VPC = 0.03$. For the contextual effect, the coverage rates are generally close to 95% when the effect size equals 0 and slightly lower when the effect size equals -0.2.

Table 3.9: 95% coverage for selected parameters based on weak convergence criteria and all converged replications

VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation latent variable between ω_B										
0.03	0.1	0.1	-0.2	99.4	14.7	25.1	99.1	29.5	1.1	29.7
0.03	0.1	0.1	0	99.2	13.5	25.1	97.8	29.1	1.8	28.4
0.03	0.1	1	-0.2	100.0	35.4	42.1	NA	56.9	NA	52.1
0.03	0.1	1	0	100.0	32.2	40.5	99.5	54.8	NA	48.3
0.03	1	0.1	-0.2	3.8	1.2	0.7	3.7	0.2	NA	0.2
0.03	1	0.1	0	5.2	2.1	1.2	3.8	1.0	NA	1.4
0.03	1	1	-0.2	7.2	4.0	0.8	NA	0.8	NA	0.8
0.03	1	1	0	7.5	3.8	1.6	5.7	1.0	NA	0.8
0.3	0.1	0.1	-0.2	100.0	57.0	NA	100.0	95.7	2.5	93.3
0.3	0.1	0.1	0	100.0	55.6	NA	100.0	97.6	2.5	95.0
0.3	0.1	1	-0.2	99.8	78.0	NA	NA	99.6	1.9	98.1
0.3	0.1	1	0	100.0	79.0	NA	100.0	99.0	2.9	98.3
0.3	1	0.1	-0.2	9.5	3.8	1.2	NA	0.4	NA	0.2
0.3	1	0.1	0	12.0	5.3	2.2	7.9	1.6	NA	1.2
0.3	1	1	-0.2	27.8	17.0	2.2	NA	1.9	NA	1.3
0.3	1	1	0	30.5	17.4	5.8	21.0	3.9	NA	3.1
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation latent variable within ω_W										
0.03	0.1	0.1	-0.2	100.0	19.4	12.8	100.0	45.9	0.0	34.2
0.03	0.1	0.1	0	100.0	18.4	13.5	100.0	44.8	0.0	34.9
0.03	0.1	1	-0.2	2.4	0.8	0.2	NA	0.0	NA	0.0
0.03	0.1	1	0	1.8	0.8	0.2	0.9	0.2	NA	0.2
0.03	1	0.1	-0.2	100.0	43.1	33.3	100.0	65.6	NA	48.7
0.03	1	0.1	0	100.0	41.6	28.7	100.0	64.6	NA	50.8
0.03	1	1	-0.2	7.6	4.4	0.0	NA	0.0	NA	0.0
0.03	1	1	0	4.6	1.6	0.4	3.6	0.0	NA	0.2
0.3	0.1	0.1	-0.2	100.0	26.0	NA	100.0	39.3	0.0	32.8
0.3	0.1	0.1	0	100.0	29.4	NA	100.0	38.8	0.0	36.2
0.3	0.1	1	-0.2	2.4	0.9	NA	NA	0.0	0.0	0.2
0.3	0.1	1	0	0.8	0.0	NA	1.0	0.0	0.0	0.0
0.3	1	0.1	-0.2	100.0	51.9	43.8	NA	71.4	NA	57.4
0.3	1	0.1	0	100.0	47.7	39.5	100.0	66.4	NA	54.5
0.3	1	1	-0.2	7.2	4.0	1.3	NA	0.2	NA	0.2
0.3	1	1	0	5.4	2.9	0.4	2.0	0.0	NA	0.0
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Standard deviation item 1 between σ^B_{y1}										
0.03	0.1	0.1	-0.2	98.0	86.2	74.7	97.6	78.0	9.1	73.0
0.03	0.1	0.1	0	97.8	86.7	75.6	97.6	78.4	6.3	73.1
0.03	0.1	1	-0.2	97.4	85.6	76.2	NA	77.4	NA	74.6
0.03	0.1	1	0	98.0	85.7	77.9	96.4	79.1	NA	75.4
0.03	1	0.1	-0.2	93.1	93.7	92.0	93.8	92.8	NA	92.8
0.03	1	0.1	0	93.4	93.2	93.1	93.4	92.4	NA	91.2
0.03	1	1	-0.2	95.2	94.2	92.5	NA	90.2	NA	90.4
0.03	1	1	0	95.4	95.2	93.1	95.1	90.8	NA	91.6
0.3	0.1	0.1	-0.2	94.3	94.9	NA	95.5	91.1	91.6	91.4
0.3	0.1	0.1	0	93.7	92.1	NA	93.4	92.4	91.9	92.2
0.3	0.1	1	-0.2	95.4	95.3	NA	NA	90.6	92.0	92.1
0.3	0.1	1	0	97.8	98.2	NA	96.7	93.7	95.5	93.8
0.3	1	0.1	-0.2	94.9	95.3	94.0	NA	93.8	NA	94.0
0.3	1	0.1	0	93.8	94.0	94.4	94.2	94.9	NA	95.0
0.3	1	1	-0.2	94.9	95.6	93.7	NA	91.3	NA	91.6
0.3	1	1	0	94.6	95.2	93.1	95.7	91.8	NA	93.0
VPC	ω_B	ω_W	ES	HNinf	HCinf	Finf	EXPinf	HNinc	HCinc	EXPinc
Contextual effect β_B - β_W										
0.03	0.1	0.1	-0.2	96.4	91.5	90.6	96.0	93.1	86.0	92.4
0.03	0.1	0.1	0	96.8	94.5	93.7	96.8	95.9	92.5	95.3
0.03	0.1	1	-0.2	56.9	44.5	39.6	NA	39.5	NA	37.2
0.03	0.1	1	0	95.4	93.4	92.5	95.0	94.2	NA	93.6
0.03	1	0.1	-0.2	95.2	94.3	93.7	94.9	93.2	NA	93.6
0.03	1	0.1	0	92.8	93.0	93.3	93.2	92.8	NA	93.0
0.03	1	1	-0.2	85.9	85.4	76.5	NA	81.1	NA	76.7
0.03	1	1	0	94.8	93.6	92.9	93.3	94.0	NA	93.0
0.3	0.1	0.1	-0.2	95.5	92.4	NA	96.2	93.5	92.7	92.9
0.3	0.1	0.1	0	96.4	94.2	NA	97.1	93.7	94.3	93.5
0.3	0.1	1	-0.2	78.8	75.6	NA	NA	68.3	64.1	68.2
0.3	0.1	1	0	97.2	96.8	NA	96.9	94.7	94.2	95.1
0.3	1	0.1	-0.2	96.0	96.2	94.5	NA	95.1	NA	94.9
0.3	1	0.1	0	96.4	96.2	95.2	96.0	95.7	NA	95.7
0.3	1	1	-0.2	90.1	90.4	84.8	NA	88.0	NA	84.1
0.3	1	1	0	96.7	96.5	95.5	96.7	95.7	NA	95.3

Note. Finc is missing due to complete nonconvergence. NA indicates that results are not available since the convergence percentage < 50% in this condition.

Results Study 3: Influence number of groups

In this study, we investigate the influence of increasing the number of groups from 20 to 50. We analyse a subset of the conditions from Study 1 and 2, with the default and correctly specified informative priors.

Convergence

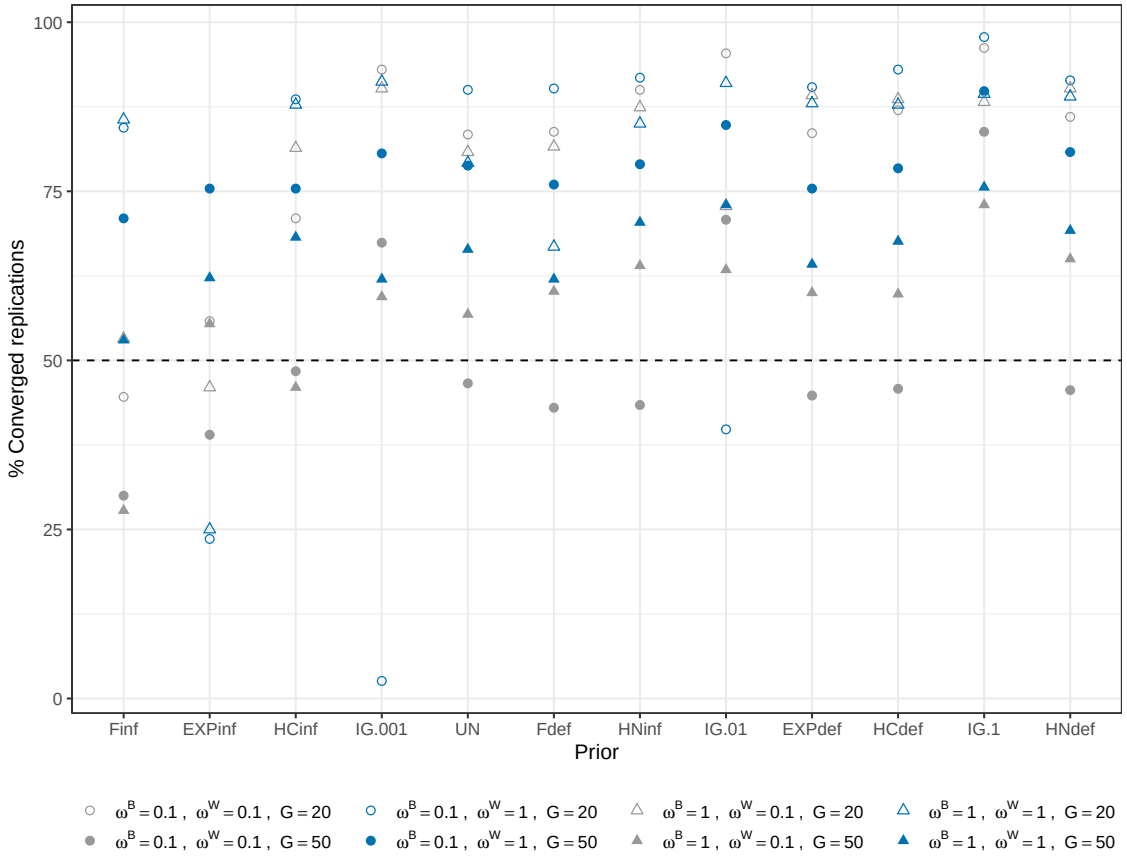

Figure 3.11: Percentages converged replications per condition according to the strict convergence criteria

Figure 3.11 shows the percentages of converged replications according to the strict criteria for each prior in the third study (filled shapes). For comparison, the convergence percentages for $G = 20$ are shown as well (empty shapes). Surprisingly, convergence is not consistently higher for $G = 50$ compared to $G = 20$. This might be due to the fact that with an increased number of groups, there are more factor scores at the between level η_j^B as well as latent group means $x_{b.j}$ to estimate which means that there are more parameters for which Rhat might become too large. Generally, for $G = 50$, convergence is above the required 50%, except for many priors in the condition with $\omega_B = \omega_W = 0.1$ (grey dots) and for two priors when $\omega_B = 1, \omega_W = 0.1$ (grey triangles).

Bias

Table 3.10 shows the relative bias with the absolute bias in brackets for the variance parameters and the parameter of interest, $\beta_B - \beta_W$, based on the posterior median estimates and strict convergence criteria. Differences between the priors are generally small, with the exception of the inverse-Gamma priors which show higher or lower biases depending on the population condition. Compared to the bias found in Study 1 and 2, there are some differences in bias, in which case the bias is generally lower for $G = 50$. However, there is no clear structure to these differences.

Table 3.10: Relative bias with absolute bias in brackets for selected parameters based on strict convergence criteria and all converged replications

ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
						Standard deviation latent variable between ω_B							
0.1	0.1	NA	-0.418 (-0.042)	0.123 (0.012)	1.361 (0.136)	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	-0.459 (-0.046)	-0.42 (-0.042)	0.074 (0.007)	1.253 (0.125)	-0.444 (-0.044)	-0.445 (-0.045)	-0.44 (-0.044)	-0.465 (-0.047)	-0.513 (-0.051)	-0.836 (-0.084)	-0.85 (-0.085)	-0.557 (-0.056)
1	0.1	-0.856 (-0.856)	-0.895 (-0.895)	-0.817 (-0.817)	-0.636 (-0.636)	-0.858 (-0.858)	-0.857 (-0.857)	-0.857 (-0.857)	-0.866 (-0.866)	-0.873 (-0.873)	NA	NA	-0.876 (-0.876)
1	1	-0.878 (-0.878)	-0.908 (-0.908)	-0.84 (-0.84)	-0.679 (-0.679)	-0.872 (-0.872)	-0.877 (-0.877)	-0.881 (-0.881)	-0.884 (-0.884)	-0.881 (-0.881)	-0.881 (-0.881)	-0.899 (-0.899)	-0.885 (-0.885)
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
						Standard deviation latent variable within ω_W							
0.1	0.1	NA	-0.305 (-0.031)	0.476 (0.048)	2.393 (0.239)	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	-0.825 (-0.825)	-0.877 (-0.877)	-0.789 (-0.789)	-0.598 (-0.598)	-0.83 (-0.83)	-0.833 (-0.833)	-0.829 (-0.829)	-0.839 (-0.839)	-0.838 (-0.838)	-0.875 (-0.875)	-0.896 (-0.896)	-0.848 (-0.848)
1	0.1	-0.059 (-0.006)	-0.293 (-0.029)	0.339 (0.034)	1.862 (0.186)	-0.134 (-0.013)	-0.155 (-0.016)	-0.142 (-0.014)	-0.208 (-0.021)	-0.358 (-0.036)	NA	NA	-0.434 (-0.043)
1	1	-0.826 (-0.826)	-0.882 (-0.882)	-0.819 (-0.819)	-0.661 (-0.661)	-0.838 (-0.838)	-0.84 (-0.84)	-0.845 (-0.845)	-0.856 (-0.856)	-0.834 (-0.834)	-0.842 (-0.842)	-0.878 (-0.878)	-0.846 (-0.846)
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
						Standard deviation item 1 between σ^B_{y1}							
0.1	0.1	NA	-0.174 (-0.058)	-0.113 (-0.038)	0.031 (0.01)	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	-0.059 (-0.022)	-0.177 (-0.067)	-0.115 (-0.043)	-0.002 (-0.001)	-0.075 (-0.028)	-0.077 (-0.029)	-0.075 (-0.028)	-0.093 (-0.035)	-0.085 (-0.032)	-0.23 (-0.087)	-0.301 (-0.114)	-0.111 (-0.042)
1	0.1	0.02 (0.021)	-0.008 (-0.008)	-0.004 (-0.004)	-0.023 (-0.025)	-0.009 (-0.009)	-0.005 (-0.005)	0.003 (0.003)	-0.004 (-0.004)	0.01 (0.011)	NA	NA	-0.001 (-0.001)
1	1	0.018 (0.019)	-0.005 (-0.005)	0.002 (0.002)	-0.008 (-0.009)	0.004 (0.004)	0.004 (0.004)	-0.001 (-0.002)	0.002 (0.002)	0.015 (0.016)	0.011 (0.011)	-0.005 (-0.006)	0.022 (0.023)
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
						Contextual effect β_B - β_W							
0.1	0.1	NA	-0.903 (0.013)	-0.789 (0.011)	-0.859 (0.012)	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	-0.988 (0.103)	-0.996 (0.104)	-0.984 (0.102)	-0.968 (0.101)	-0.984 (0.102)	-0.989 (0.103)	-0.978 (0.102)	-0.977 (0.102)	-0.988 (0.103)	-0.986 (0.103)	-0.991 (0.103)	-0.985 (0.102)
1	0.1	-0.686 (0.01)	-0.72 (0.01)	-0.629 (0.009)	-0.696 (0.01)	-0.824 (0.012)	-0.708 (0.01)	-0.947 (0.013)	-0.989 (0.014)	-0.765 (0.011)	NA	NA	-0.782 (0.011)
1	1	-1.027 (0.107)	-1.017 (0.106)	-1.026 (0.107)	-1.023 (0.106)	-1.03 (0.107)	-1.041 (0.108)	-1.058 (0.11)	-1.031 (0.107)	-1.038 (0.108)	-1.013 (0.105)	-0.994 (0.103)	-1.029 (0.107)

Note. In Study 3, only those conditions are considered in which $VPC = 0.03$ and $ES = -0.2$. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Mean squared error (MSE)

The MSE for the variance parameters and the contextual effect, $\beta_B - \beta_W$, is shown in Table 3.11. Since these values do not differ substantially from those obtained based on 20 groups in Study 1 and 2, we will not discuss them further.

Table 3.11: Mean squared error (MSE) for selected parameters based on strict convergence criteria and all converged replications

ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation latent variable between ω_B													
0.1	0.1	NA	0.002	0.001	0.019	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	0.003	0.002	0.000	0.016	0.003	0.003	0.003	0.003	0.003	0.007	0.008	0.003
1	0.1	0.738	0.802	0.670	0.410	0.741	0.740	0.738	0.754	0.766	NA	NA	0.770
1	1	0.777	0.827	0.708	0.465	0.765	0.773	0.780	0.784	0.779	0.780	0.812	0.788
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation latent variable within ω_W													
0.1	0.1	NA	0.001	0.003	0.061	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	0.688	0.772	0.626	0.364	0.696	0.699	0.694	0.710	0.707	0.769	0.805	0.724
1	0.1	0.002	0.001	0.002	0.037	0.002	0.002	0.002	0.002	0.002	NA	NA	0.002
1	1	0.689	0.780	0.673	0.442	0.708	0.711	0.719	0.737	0.702	0.715	0.774	0.720
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation item 1 between σ^B_{y1}													
0.1	0.1	NA	0.018	0.010	0.004	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	0.013	0.020	0.012	0.006	0.013	0.013	0.013	0.014	0.013	0.022	0.037	0.014
1	0.1	0.022	0.022	0.021	0.023	0.018	0.021	0.020	0.018	0.018	NA	NA	0.019
1	1	0.025	0.022	0.023	0.028	0.023	0.023	0.022	0.022	0.022	0.021	0.022	0.024
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Contextual effect β_B - β_W													
0.1	0.1	NA	0.001	0.002	0.006	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	0.011	0.011	0.012	0.014	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
1	0.1	0.004	0.002	0.006	0.017	0.003	0.003	0.004	0.003	0.003	NA	NA	0.002
1	1	0.014	0.013	0.015	0.021	0.014	0.015	0.015	0.014	0.014	0.013	0.013	0.014

Note. In Study 3, only those conditions are considered in which $VPC = 0.03$ and $ES = -0.2$. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Coverage

The general structure of the coverage rates based on the 95% credibility intervals (Table 3.12) is very similar to the values found in Study 1 and 2. However, if the coverage rates differ, the coverage is generally lower in Study 3. One substantial difference across all priors is that the coverage for the contextual effect, β_B - β_W is much lower when $\omega_W = 1$ in Study 3 compared to Study 1 and 2.

Table 3.12: 95% coverage for selected parameters based on strict convergence criteria and all converged replications

ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation latent variable between ω_B													
0.1	0.1	NA	98.5	100.0	0.0	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	97.7	97.3	100.0	0.0	98.0	97.7	97.1	97.9	96.5	34.5	40.3	93.9
1	0.1	2.1	0.0	1.6	9.6	1.2	1.0	2.3	1.3	0.6	NA	NA	0.0
1	1	2.1	0.3	1.1	4.2	1.4	0.6	0.3	1.2	1.4	1.2	0.8	1.3
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation latent variable within ω_W													
0.1	0.1	NA	100.0	99.2	0.0	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	2.3	0.5	0.9	12.0	1.2	1.5	1.6	0.5	0.8	0.3	0.0	0.5
1	0.1	100.0	100.0	100.0	0.0	100.0	100.0	100.0	100.0	100.0	NA	NA	100.0
1	1	3.6	0.3	0.5	5.8	2.0	1.2	1.0	0.9	3.4	1.8	0.0	2.3
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Standard deviation item 1 between σ^B_{y1}													
0.1	0.1	NA	89.6	96.6	99.3	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	93.7	87.6	92.2	98.4	93.3	93.4	93.2	92.6	92.7	86.7	81.4	91.8
1	0.1	95.8	93.9	95.3	95.3	96.3	95.7	97.3	96.0	97.2	NA	NA	96.0
1	1	94.9	96.8	94.5	93.9	94.2	95.3	96.5	95.6	96.0	95.6	95.8	96.8
ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef	HNinf	HCinf	Finf	EXPinf
Contextual effect β_B - β_W													
0.1	0.1	NA	92.9	96.3	98.6	NA	NA	NA	NA	NA	NA	NA	NA
0.1	1	31.7	18.9	46.0	79.1	33.7	31.9	35.8	31.3	28.6	16.4	12.1	25.2
1	0.1	92.3	90.6	92.1	93.7	92.6	93.6	92.0	93.0	92.2	NA	NA	93.5
1	1	63.0	44.5	61.4	83.1	61.6	58.0	58.1	57.3	62.2	61.6	50.9	57.9

Note. In Study 3, only those conditions are considered in which $VPC = 0.03$ and $ES = -0.2$. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Summary of the simulation studies

With these three simulation studies, we aimed to obtain more insight into the performance of the various priors for random effects variances in MLSEM. Here, we summarise the results across the three studies. Generally, the default robust priors and the uniform prior show sufficient convergence, even with only 20 groups. The inverse-Gamma priors and informative robust priors show more convergence issues. In certain population conditions (i.e., $\omega_B = \omega_W = 0.1$), convergence percentages are lower for some priors when $G = 50$ compared to $G = 20$. This might be due to the strict convergence criteria used. Specifically, only one divergent transition, or one parameter with an $\hat{R}$ value larger than 1.05 result in a replication being flagged as non-converged. As the group size and thus the model grows, this might lead to the higher nonconvergence percentages.

The varying results for the different inverse-Gamma specifications replicate previous findings showing that these priors are highly sensitive to the choice of the hyperparameters. The uniform and robust priors show substantial negative bias and much too low coverage rates for both ω_B and ω_W when the population values for these parameters are equal to 1. This indicates that, regardless of their heavy tails, the robust priors shrink large variance parameters too heavily towards zero and are not sufficiently robust for population values equal to 1. Moreover, these problems are not remedied by specifying informative prior distributions. Coverage rates for the latent variable standard deviations do improve when ω_B and ω_W are equal to 0.1, as does the bias for ω_W when its population value equals 0.1. Interestingly, the standard deviation of the items at the between level (σ_y^B) is much less sensitive to the prior specification.

Although the priors we investigated are specified on the random effects variances, researchers using MLSEM are generally interested in the contextual effect $\beta_B - \beta_W$. All of the investigated priors result in an underestimation of the contextual effect, regardless of condition. Coverage of the contextual effect is generally close to the nominal 95% when the population effect size equals 0 or when it equals -0.2 in combination with $\omega_W = 0.1$. Otherwise, coverage rates are too low. Interestingly, for $G = 50$ the coverage rates are substantially lower when $ES = -0.2$ and $\omega_W = 1$ compared to $G = 20$. This can be explained by the fact that, in general, the credibility intervals are smaller for $G = 50$ compared to $G = 20$. However, if the parameter estimate is biased, this smaller credibility interval will not contain the true value more often compared to the wider credibility interval, leading to worse coverage rates. Finally, the power to detect a small contextual effect of -0.2 is much too low across the board. This is not surprising given the small number of groups used.

3.7 Discussion

The popularity of Bayesian MLSEM has inspired various investigations into the required number of groups (see e.g., Depaoli & Clifton, 2015; Helm, 2018; Hox et al., 2012). However, these studies mainly rely on traditional default prior specifications, that have been proven to be unreliable, or correctly specified informative priors that are not feasible in practice. The goal of this study was to investigate more robust prior distributions for the random effects variances in the context of MLSEM. In order to do so, we have conducted three simulation studies and applied the priors to the PISA data to estimate the BFLPE.

Overall, the differences between the prior distributions were smaller than expected. The main differences in results in the simulation studies appear to arise

due to the population conditions rather than the prior distribution. This might be explained by the fact that the conditions considered in our simulations were not extreme enough to result in substantial differences between the priors. Although we tried to include a variety of realistic simulation conditions, limitations always remain due to time constraints. Specifically, we only considered balanced designs with a within-group sample size of 20. Future research should further investigate non-balanced designs and vary the within-group sample size. Additionally, for a smaller number of groups (e.g., 10), we would expect more differences between the priors. However, the question arises how useful such a situation would be in practice, especially given the low power we encountered for 20 and even 50 groups. One prior that did show substantially different results is the traditional inverse-Gamma prior, which has once again shown to be highly sensitive to the specific choice of the hyperparameters. Unfortunately, the informative priors resulted in low convergence percentages which complicated a thorough investigation and comparison of their performance. As expected, the bias was generally lower for the correctly specified priors compared to the incorrectly specified priors, but the exact value depended on the population condition. Thus, in certain conditions, there was some evidence that these priors are robust against misspecification. Future research should further investigate these and other informative specifications to fully assess their robustness against misspecification.

One limitation of the simulation study lies in the fact that we removed the non-converged replications from the results. If the non-convergence in these removed replications is due to particular sample values this approach might bias the simulations. We tried to partially remedy this problem by also considering weaker convergence criteria, which resulted in less non-converged replications but generally did not show substantial differences in terms of bias, MSE, and coverage. Still, non-convergence and how to deal with it remains an issue in any simulation study investigating extreme conditions in terms of sample size or population values. Compared to other MCMC algorithms, the Hamiltonian Monte Carlo algorithm used in Stan offers more convergence criteria. An advantage is that convergence can be assessed more accurately, however, it will also flag a replication as non-converged more easily. Especially for large models such as the model considered here, a few divergent transitions can occur rapidly resulting in a replication that is flagged as non-converged. However, in practice, removing all non-converged replications based on only a few divergent transitions might result in more bias than when we do not remove these replications. It can therefore be useful to compare the results obtained using various convergence criteria.

Throughout this paper, we have focused on the prior for the random effects variance or standard deviation. In regular multilevel models, several authors have

proposed to specify priors on the intraclass-correlation coefficient, which leads to an implied prior on the random effects variance (see e.g., Daniels, 1999; Gustafson, Hossain, & Macnab, 2006; Mulder & Fox, 2018; Natarajan & Kass, 2000). Although we have considered specifying a prior directly on the VPC, this approach is not straightforward in MLSEM. Specifically, since the VPC depends on multiple other model parameters (see Equation (3.7)), a choice needs to be made for which of those model parameters a prior is directly specified and for which of the model parameters the prior is implied. Future research is needed to investigate these choices and their consequences.

Bayesian MLSEM offers a powerful modeling framework to incorporate both measurement and sampling error within one model. However, the results presented in this paper indicate that some caution is warranted in applying these models. When the sample size is small, Bayesian estimation does not necessarily perform well (Smid, McNeish, Miočević, & van de Schoot, 2019). In order for Bayesian estimation to outperform classical estimation methods, the prior distribution needs to be well thought out. This is especially the case for the complex MLSEM considered in this paper: with only 20 groups, it is crucial for the prior distribution to add reasonable prior information to the analysis. More research is needed to investigate how we can achieve this.

Chapter 4

Shrinkage priors for Bayesian penalized regression.

Based on van Erp, S., Oberski, D.L., Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31-50. doi:10.1016/j.jmp.2018.12.004

Abstract

In linear regression problems with many predictors, penalized regression techniques are often used to guard against overfitting and to select variables relevant for predicting an outcome variable. Recently, Bayesian penalization is becoming increasingly popular in which the prior distribution performs a function similar to that of the penalty term in classical penalization. Specifically, the so-called *shrinkage priors* in Bayesian penalization aim to shrink small effects to zero while maintaining true large effects. Compared to classical penalization techniques, Bayesian penalization techniques perform similarly or sometimes even better, and they offer additional advantages such as readily available uncertainty estimates, automatic estimation of the penalty parameter, and more flexibility in terms of penalties that can be considered. However, many different shrinkage priors exist and the available, often quite technical, literature primarily focuses on presenting one shrinkage prior and often provides comparisons with only one or two other shrinkage priors. This can make it difficult for researchers to navigate through the many prior options and choose a shrinkage prior for the problem at hand. Therefore, the aim of this paper is to provide a comprehensive overview of the literature on Bayesian penalization. We provide a theoretical and conceptual comparison of nine different shrinkage priors and parametrize the priors, if possible, in terms of scale mixture of normal distributions to facilitate comparisons. We illustrate different characteristics and behaviors of the shrinkage priors and compare their performance in terms of prediction and variable selection in a simulation study. Additionally, we provide two empirical examples to illustrate the application of Bayesian penalization. Finally, an R package `bayesreg` is available online (`https://github.com/sara-vanerp/bayesreg`) which allows researchers to perform Bayesian penalized regression with novel shrinkage priors in an easy manner.

Keywords: Bayesian, Shrinkage Priors, Penalization, Empirical Bayes, Regression.

4.1 Introduction

Regression analysis is one of the main statistical techniques often used in the field of psychology to determine the effect of a set of predictors on an outcome variable. The number of predictors is often large, especially in the current "Age of Big Data". For example, the Kavli HUMAN project (Azmak et al., 2015) aims to collect longitudinal data on all aspects of human life for 10,000 individuals. Measurements include psychological assessments (e.g., personality, IQ), health assessments (e.g., genome sequencing, brain activity scanning), social network assessment, and variables related to education, employment, and financial status, resulting in an extremely large set of variables. Furthermore, personal tracking devices allow the collection of large amounts of data on various topics, including for example mood, in a longitudinal manner (Fawcett, 2015). The problem with regular regression techniques such as ordinary least squares (OLS) is that they quickly lead to overfitting as the ratio of predictor variables to observations increases (see for example, McNeish, 2015, for an overview of the problems with OLS).

Penalized regression is a statistical technique widely used to guard against overfitting in the case of many predictors. Penalized regression techniques have the ability to select variables out of a large set of variables that are relevant for predicting some outcome. Therefore, a popular setting for penalized regression is in high-dimensional data, where the number of predictors p is larger than the sample size n. Furthermore, in settings where the number of predictors p is smaller than the sample size n (but still relatively large), penalized regression can offer advantages in terms of avoiding overfitting and achieving model parsimony compared to traditional variable selection methods such as null-hypothesis testing or stepwise selection methods (Derksen & Keselman, 1992; Tibshirani, 1996). The central idea of penalized regression approaches is to add a penalty term to the minimization of the sum of squared residuals, with the goal of shrinking small coefficients towards zero while leaving large coefficients large, i.e.,

$$\underset{\beta_0,\,\boldsymbol{\beta}}{\text{minimize}} \left\{ \frac{1}{2n} ||\boldsymbol{y} - \beta_0 \mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_c ||\boldsymbol{\beta}||_q \right\}, \tag{4.1}$$
$$\text{where } ||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{\frac{1}{q}},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is an n-dimensional vector containing the observations on the outcome variable, β_0 reflects the intercept, $\mathbf{1}$ is an n-dimensional vector of ones, $\boldsymbol{X}$ is an $(n \times p)$ matrix of the observed scores on the p predictor variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional parameter vector of regression coefficients. λ_c

reflects the penalty parameter, with large values resulting in more shrinkage towards zero while $\lambda_c = 0$ leads to the ordinary least squares solution. The choice of q determines the type of penalty induced, for example, $q = 1$ results in the well-known least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) solution and $q = 2$ results in the ridge solution (Hoerl & Kennard, 1970). We refer to Hastie, Tibshirani, and Wainwright (2015) for a comprehensive introduction and overview of various penalized regression methods in a frequentist framework.

It is well known that many solutions to the penalized minimization problem in Equation (4.1) can also be obtained in the Bayesian framework by using a specific prior combined with the posterior mode estimate, which has been shown to perform similar to or better than their classical counterparts (Hans, 2009; Kyung, Gill, Ghosh, & Casella, 2010; Li & Lin, 2010).[1] Adopting a Bayesian perspective on penalized regression offers several advantages. First, penalization fits naturally in a Bayesian framework since a prior distribution is needed anyway and shrinkage towards zero can be straightforwardly achieved by choosing a specific parametric form for the prior. Second, parameter uncertainty and standard errors follow naturally from the posterior standard deviations. As shown by Kyung et al. (2010) classical penalized regression procedures can result in estimated standard errors that suffer from multiple problems, such as variances estimated to be 0 (in the case of sandwich estimates), and unstable or poorly performing variance estimates (in the case of bootstrap estimates). Third, with Bayesian penalization it is possible to estimate the penalty parameter(s) λ simultaneously with the model parameters in a single step. This is especially advantageous when there are multiple penalty parameters (e.g., in the elastic net; Zou & Hastie, 2005), since sequential cross-validation procedures to determine multiple penalty parameters induce too much shrinkage (i.e., the double shrinkage problem; see e.g., Zou & Hastie, 2005). Fourth, Bayesian penalization relies on Markov Chain Monte Carlo (MCMC) sampling rather than optimization, which provides more flexibility in the sense that priors that would correspond to non-convex penalties (i.e., $q < 1$ in (4.1)) are easier to implement. Non-convex penalties would result in multiple modes, making them difficult to implement in an optimization framework. The cost of the flexibility of MCMC, however, is that it requires more computation time compared to standard optimization procedures. Finally, Bayesian estimates have an intuitive interpretation. For example, a 95% Bayesian credibility interval can simply be interpreted as the interval in which the true value lies with 95% probability (e.g., Berger, 2006).

Due to these advantages, Bayesian penalization is becoming increasingly popular in the literature (see e.g., Alhamzawi, Yu, & Benoit, 2012; Andersen, Vehtari,

[1] Note that from a Bayesian perspective, however, there is no theoretical justification for reporting the posterior mode estimate (Tibshirani, 2011).

Winther, & Hansen, 2017; Armagan, Dunson, & Lee, 2013; Bae & Mallick, 2004; Bhadra, Datta, Polson, & Willard, 2017a; Bhattacharya, Pati, Pillai, & Dunson, 2012; Bornn, Gottardo, & Doucet, 2010; Caron & Doucet, 2008; Carvalho, Polson, & Scott, 2010; Feng, Wang, Lu, & Song, 2017; Griffin & Brown, 2017; Hans, 2009; Ishwaran & Rao, 2005; Lu, Chow, & Loken, 2016; Peltola, Havulinna, Salomaa, & Vehtari, 2014; Polson & Scott, 2011; Roy & Chakraborty, 2016; Zhao, Gao, Mukherjee, & Engelhardt, 2016). An active area of research investigates theoretical properties of priors for Bayesian penalization, such as the Bayesian lasso prior (for a recent overview, see Bhadra, Datta, Polson, & Willard, 2017b). In addition to the Bayesian counterparts of classical penalized regression solutions, many other priors have been proposed that have desirable properties in terms of prediction and variable selection. However, the extensive (and often technical) literature and subtle differences between the priors can make it difficult for researchers to navigate the options and make sensible choices for the problem at hand. Therefore, the aim of this paper is to provide a comprehensive overview of the priors that have been proposed for penalization in (sparse) regression. We use the term *shrinkage priors* to emphasize that these priors aim to shrink small effects towards zero. We place the shrinkage priors in a general framework of scale mixtures of normal distributions to emphasize the similarities and differences between the priors. By providing insight in the characteristics and behaviors of the priors, we aid researchers in choosing a prior for their specific problem. Additionally, we present a straightforward method to obtain empirical Bayes (EB) priors for Bayesian penalization. We conduct a simulation study to compare the performance of the priors in terms of prediction and variable selection in a linear regression model, and provide two empirical examples to further illustrate the Bayesian penalization methods. Finally, the shrinkage priors have been implemented in the R package `bayesreg`, available from `https://github.com/sara-vanerp/bayesreg`, to allow general utilization.

The remainder of this paper is organized as follows: Section 2 introduces Bayesian penalized regression. A theoretical overview of the different shrinkage priors can be found in Section 3. Further insights into the priors is provided through illustrations in Section 4 and the priors are compared in a simulation study in Section 5. Section 6 presents the empirical applications, followed by a discussion in Section 7.

4.2 Bayesian penalized regression

The likelihood for the linear regression model is given by:

$$y_i|\beta_0, \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2), \tag{4.2}$$

where β_0 represents the intercept, β_j the regression coefficient for predictor j, and σ^2 is the residual variance.

In a Bayesian analysis, a prior distribution is specified for each parameter in the model, e.g., $p(\beta_0, \boldsymbol{\beta}, \sigma^2, \lambda) = p(\beta_0)p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)p(\lambda)$. Note that the prior for $\boldsymbol{\beta}$ is conditioned on the residual variance σ^2, as well as on λ. The conditioning on σ^2 is necessary in certain cases to obtain a unimodal posterior (Park & Casella, 2008). In Bayesian penalized regression, λ is a parameter in the prior (i.e., a hyperparameter) but has a similar role as the penalty parameter in classical penalized regression. Since this penalty parameter λ is used to penalize the regression coefficient, it only appears in the prior for $\boldsymbol{\beta}$. Throughout this paper we will focus on priors for the regression coefficients $\beta_1, \ldots, \beta_j$ and we will assume noninformative improper priors for the nuisance parameters, specifically, $p(\beta_0) = 1$ and a uniform prior on $\log(\sigma^2)$, i.e., $p(\sigma^2) = \sigma^{-2}$. Please note that these priors are chosen as noninformative choices for the linear regression model considered in this paper. However, other choices (including informative priors when prior information is available) are possible and might be preferred in other applications. We refer the reader to van Erp et al. (2018) for general recommendations on specifying prior distributions. We generally assume that the priors for the regression coefficients are independent, unless stated otherwise.

The prior distribution is then multiplied by the likelihood of the data to obtain the posterior distribution, i.e.,

$$p(\beta_0, \boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}, X) \propto p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2)p(\beta_0)p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)p(\lambda). \tag{4.3}$$

Here, the normalizing constant is not included such that the right-hand side is proportional to the posterior. The only difference with the unpenalized problem (e.g., Bayesian linear regression) is the introduction of the penalty parameter λ. As a result of the shrinkage prior the posterior in (4.3) is generally more concentrated, or "shrunk towards", zero in comparison to the likelihood of the model.

An important choice is how to specify the penalty parameter λ. There are different possibilities for this.

1. *Full Bayes.* Treat λ as an unknown model parameter for which a prior needs to be specified. Typically, a vague prior $p(\lambda)$ is specified for λ. Due to its

similarity with multilevel (or hierarchical) modeling, full Bayes (FB) is also known as "hierarchical Bayes" (see e.g., Wolpert & Strauss, 1996). This results in a fully Bayesian solution that incorporates the uncertainty about λ. The advantage of this approach is that the model can be estimated in one step. Throughout this paper, we will consider the half-Cauchy prior on λ, i.e., $\lambda \sim$ half-Cauchy(0, 1), which is a robust alternative and a popular prior distribution in the Bayesian literature (see e.g., Gelman, 2006; Mulder & Pericchi, 2018; Polson & Scott, 2012).

2. *Empirical Bayes.* Empirical Bayes (EB) methods, also known as the "evidence" procedure (see e.g., Wolpert & Strauss, 1996), first estimate the penalty parameter λ from the data and then plug in this EB estimate for λ in the model (see van de Wiel, Beest, & Münch, 2017, for an overview of EB methodology in high-dimensional data). The resulting prior is called an EB prior. Since an EB estimate is used for λ, the EB approach does not require the specification of a prior $p(\lambda)$ as in the FB approach. Since the exact choice of this prior can sometimes have a serious effect on the Bayesian estimates (Roy & Chakraborty, 2016), the EB approach would avoid sensitivity of the results to the exact choice of the prior $p(\lambda)$, while keeping the advantages of the Bayesian approach.

   Empirical Bayes is a two-step approach: first, the empirical Bayes choice for λ needs to be determined; second, the model is fitted using the EB prior. In order to obtain an EB estimate for λ, we need to find the solution that maximizes the marginal likelihood[2], i.e.,

$$\lambda^{EB} = \arg\max p(\boldsymbol{y}|\lambda). \tag{4.4}$$

   To obtain λ^{EB}, first note that the marginal likelihood is the product of the likelihood and prior integrated over the model parameters, i.e.,

$$p(\boldsymbol{y}|\lambda) = \iiint p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2) p(\beta_0) p(\boldsymbol{\beta}|\sigma^2, \lambda) p(\sigma^2)\, d\beta_0\, d\boldsymbol{\beta}\, d\sigma^2. \tag{4.5}$$

   Instead of directly optimizing, we achieve (4.4) by sampling from the posterior with a noninformative prior for λ.[3] The EB estimate λ^{EB} is the mode of the marginal posterior for λ, i.e., $p(\lambda|\boldsymbol{y})$. This corresponds to the maximum of the marginal likelihood $p(\boldsymbol{y}|\lambda)$ because of the noninformative prior for λ.

[2]The marginal likelihood quantifies the probability of observing the data given the model. Therefore, plugging in the EB estimate for λ will result in a prior that predicts the observed data best.

[3]Specifically, we use $\lambda \sim$ half-Cauchy$(0, 10000)$ to ensure a stable MCMC sampler.

3. *Cross-validation.* For cross-validation (CV), the data is split into a training, validation, and test set. The goal is to find a value for λ which results in a model that is accurate in predicting new data, i.e., a generalizable model that captures the signal in the data, but does not overfit (Hastie et al., 2015). To find this value for λ, a range of values is considered, using the training data $\boldsymbol{y}^{train}$ to fit all models with the different λ values. Next, each resulting model is used to predict the responses in the validation set $\boldsymbol{y}^{val}$. The value for λ that minimizes some loss function is selected, i.e.,

$$\lambda^{CV} = \arg\min L(\boldsymbol{y}^{train}, \boldsymbol{y}^{val}). \tag{4.6}$$

Given that the loss function is the negative of the log likelihood, this is equivalent to:

$$\lambda^{CV} = \arg\max p(\boldsymbol{y}^{val}|\boldsymbol{y}^{train}, \lambda). \tag{4.7}$$

Finally, λ^{CV} is used to fit the model on the test set. Generally, the prediction mean squared error (PMSE) is used to determine λ^{CV}, which corresponds to a quadratic loss function.

In practice k-fold cross-validation is often used. k-fold cross-validation is a specific implementation of cross-validation in which the data is split in only a training and a test set. The training set is split in K parts (usually $K = 5$ or $K = 10$) and the range of λ values is applied K times on $K - 1$ parts of the training set, each time with a different part as validation set. The K estimates of the PMSE are then averaged and a standard error is computed.

Frequentist penalization approaches often rely on cross-validation. In the Bayesian literature, full and empirical Bayes are often employed, although cross-validation is also possible in a Bayesian approach (see for example the `loo` package in R; Vehtari, Gabry, Yao, & Gelman, 2018). The intuition behind empirical Bayes and cross-validation is similar: empirical Bayes aims to choose the value for λ that is best in predicting the full data set, while cross-validation aims to choose the value for λ that is best in predicting the validation set given a training set. A possible disadvantage of empirical Bayes and cross-validation is that the (marginal) likelihood can be flat or multimodal when there are multiple penalty parameters (van de Wiel et al., 2017).[4]

[4]In the initial empirical Bayes approach we used a uniform prior for λ and this problem became

Throughout this paper, we will focus on the full and empirical Bayes approach to determine λ, and only consider cross-validation for the frequentist penalization methods we will compare the priors to.

4.3 Overview shrinkage priors

In this section we will give a general overview of shrinkage priors that have been proposed in the literature. Given the extensive number of shrinkage priors that has been investigated, we will limit the overview to priors that are related to well-known classical penalization methods and shrinkage priors that are popular in the Bayesian literature. Given that most shrinkage priors fall into these categories, the resulting overview, while not exhaustive, is intended to be comprehensive and will help researchers to navigate through this literature. In total, we will discuss nine different shrinkage priors.

Many continuous, unimodal, and symmetric distributions can be parametrized as a scale mixture of normals meaning that the distribution is rewritten as a normal distribution (i.e., Normal(μ, σ^2)) where the scale parameter is given a mixing density $h(\sigma^2)$ (see e.g., West, 1987). Where possible, we will present the different priors in a common framework by providing the scale mixture of normals formulation for each prior. Using this formulation, the theoretical differences and similarities between the priors become more clear and, additionally, the scale mixture of normals formulation can be computationally more efficient.

We will now describe each prior in turn. The densities for several of the shrinkage priors are presented in Table 4.1 and plotted in Figure 4.1. We will consider a full and empirical Bayes approach to obtain the penalty parameter λ (see Section 4.2) for all shrinkage priors, unless stated otherwise. For the full Bayesian approach, we will consider standard half-Cauchy priors for the penalty parameters as a robust default prior choice. We have included this choice for the prior on λ in the descriptions below, but note that other choices are possible as well.

evident through non-convergence of the sampler or extreme estimates for λ^{EB}. The problem was solved by using the half-Cauchy prior instead.

Table 4.1: Conditional prior densities for the regression coefficients $\boldsymbol{\beta}$ implied by the various shrinkage priors and references for each shrinkage prior.

Shrinkage prior	Conditional prior density $p(\beta_j \lvert \lambda, \ldots)$	Reference
Ridge	$p(\beta_j \lvert \sigma^2, \lambda) = \sqrt{\frac{\lambda}{2\pi\sigma^2}} \exp\left\{-\frac{\lambda\beta_j^2}{2\sigma^2}\right\}$	Hsiang (1975)
Local Student's t	$p(\beta_j \lvert \sigma^2, \lambda) = \frac{\sigma^2}{\pi\lambda}\left(1 + (\frac{\sigma^2}{\lambda\beta_j})^2\right)$	Griffin and Brown (2005); Meuwissen, Hayes, and Goddard (2001)
Lasso	$p(\beta_j \lvert \sigma^2, \lambda) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{\frac{-\lambda\lvert\beta_j\rvert}{\sqrt{\sigma^2}}\right\}$	Park and Casella (2008)
Elastic net	$p(\beta_j \lvert \sigma^2, \lambda_1, \lambda_2) = C \exp\left\{-\frac{1}{2\sigma^2}(\lambda_1\lvert\beta_j\rvert + \lambda_2\beta_j^2)\right\}$	Li and Lin (2010)
Group lasso	$p(\beta_j \lvert \sigma, \lambda) = C \exp\left\{-\frac{\lambda}{\sigma}\sum_{g=1}^{G} \lvert\lvert\boldsymbol{\beta}_g\rvert\rvert\right\}$	Kyung et al. (2010)
Hyperlasso	$p(\beta_j \lvert \lambda) = \lambda(2\pi)^{\frac{1}{2}}\left[1 - \frac{(\lambda\lvert\beta_j\rvert)\{1-\Phi(\lambda\lvert\beta_j\rvert)\}}{\phi(\lambda\lvert\beta_j\rvert)}\right]$	Griffin and Brown (2011)
Horseshoe	Not analytically tractable	Carvalho et al. (2010)
Discrete normal mixture	$p(\beta_j \lvert \gamma_j, \phi_j^2) = (1-\gamma_j)\left(\frac{1}{\sqrt{2\pi\phi_j^2}} \exp\left\{-\frac{\beta_j^2}{2\phi_j^2}\right\}\right) + \gamma_j\left(\frac{1}{\pi(1+\beta_j^2)}\right)$	George and McCulloch (1993); Mitchell and Beauchamp (1988)

Note. C denotes a normalization constant. $\Phi()$ and $\phi()$ in the hyperlasso are the cumulative density function and the probability density function of the standard normal distribution.

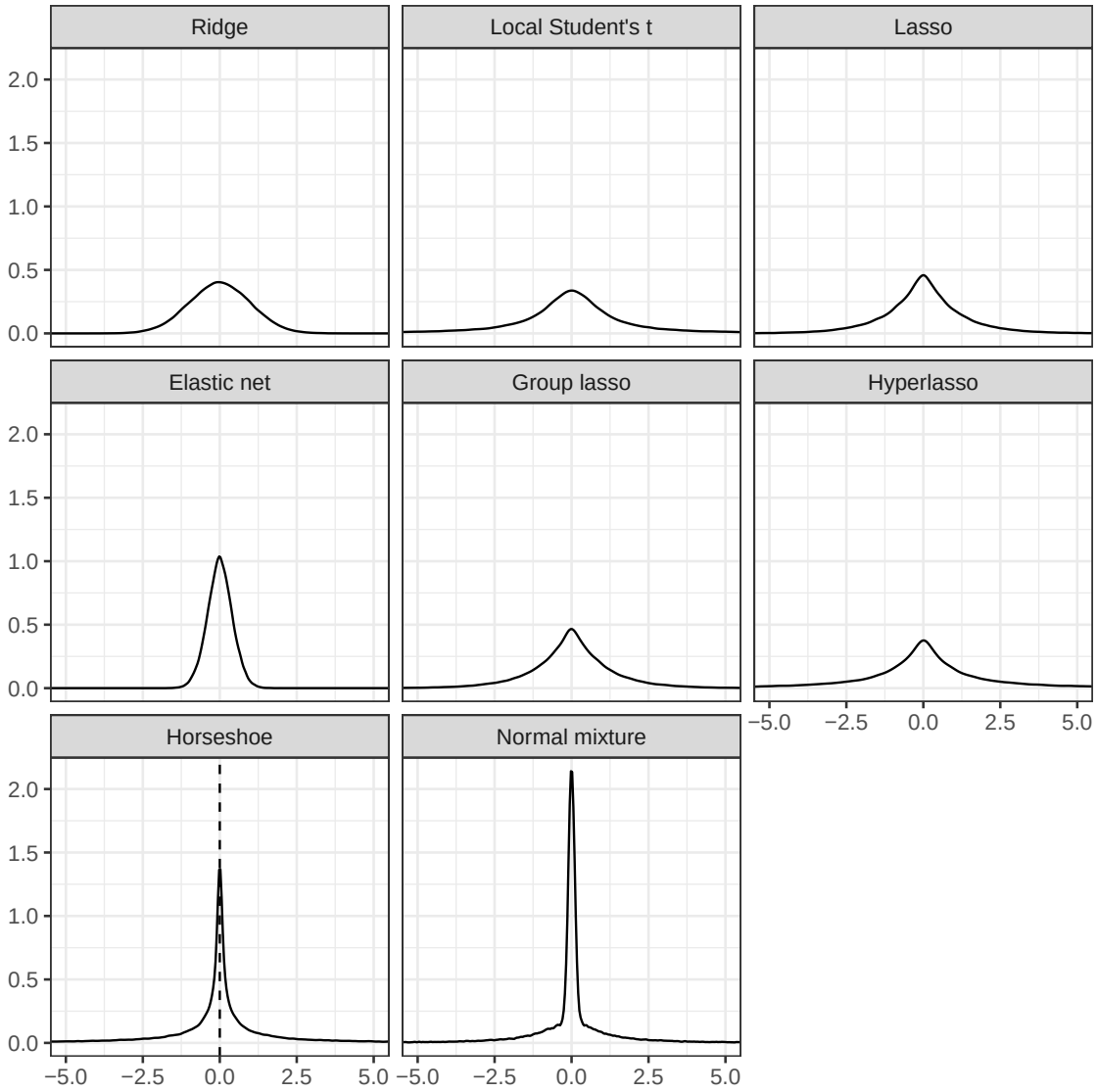

Figure 4.1: Densities of the shrinkage priors

Ridge

The ridge prior corresponds to normal priors centered around 0 on the regression coefficients, i.e., (see e.g., Hsiang, 1975)

$$\begin{aligned} \beta_j|\lambda,\sigma^2 &\sim \text{Normal}(0, \frac{\sigma^2}{\lambda}), \text{ for } j = 1,\ldots,p. \\ \lambda &\sim \text{half-Cauchy}(0,1) \end{aligned} \tag{4.8}$$

The posterior mean estimates under this prior will correspond to estimates obtained using the ridge penalty or l_2 norm, i.e., $q = 2$ in Equation (4.1) (Hoerl & Kennard, 1970). The penalty parameter λ determines the amount of shrinkage, with larger values resulting in smaller prior variation and thus more shrinkage of the coefficients

towards zero.

Local Student's t

We can extend the ridge prior in Equation 4.8 by making the prior variances predictor-specific, thereby allowing for more variation, i.e.,

$$\begin{aligned} \beta_j|\tau_j^2 &\sim \text{Normal}(0, \sigma^2\tau_j^2) \\ \tau_j^2|\nu, \lambda &\sim \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu}{2\lambda}), \text{ for } j = 1, \ldots, p, \\ \lambda &\sim \text{half-Cauchy}(0, 1) \end{aligned} \tag{4.9}$$

When integrating τ_j^2 out, the following conditional prior distribution for the regression coefficients is obtained:

$$\beta_j|\nu, \lambda, \sigma^2 \sim \text{Student}(\nu, 0, \frac{\sigma^2}{\lambda}), \tag{4.10}$$

where $\text{Student}(\nu, 0, \frac{\sigma^2}{\lambda})$ denotes a non-standardized Student's t distribution centered around 0 with ν degrees of freedom and scale parameter $\frac{\sigma^2}{\lambda}$. A smaller value for ν results in a distribution with heavier tails, with $\nu = 1$ implying a Cauchy prior for β_j. Larger (smaller) values for λ result in more (less) shrinkage towards m. This prior has been considered, among others, by Griffin and Brown (2005) and Meuwissen et al. (2001). Compared to the ridge prior in (4.8), the local Student's t prior has heavier tails. Throughout this paper, we will consider $\nu = 1$, such that the prior has Cauchy-like tails.

Lasso

The Bayesian counterpart of the lasso penalty was first proposed by Park and Casella (2008). The Bayesian lasso can be obtained as a scale mixture of normals with an exponential mixing density, i.e.,

$$\begin{aligned} \beta_j|\tau_j^2, \sigma^2 &\sim \text{Normal}(0, \sigma^2\tau_j^2) \\ \tau_j^2|\lambda^2 &\sim \text{Exponential}(\frac{\lambda^2}{2}), \text{ for } j = 1, \ldots, p, \\ \lambda &\sim \text{half-Cauchy}(0, 1) \end{aligned} \tag{4.11}$$

Integrating τ_j^2 out results in double-exponential or Laplace priors on the regression coefficients, i.e.,

$$\beta_j|\lambda,\sigma \sim \text{Double-exponential}(0, \frac{\sigma}{\lambda}), \text{ for } j = 1, \ldots, p. \tag{4.12}$$

With this prior, the posterior mode estimates are similar to estimates obtained under the lasso penalty or l_1 norm, i.e., $q = 1$ in Equation (4.1) (Tibshirani, 1996). In addition to the overall shrinkage parameter λ, the lasso prior has an additional predictor-specific shrinkage parameter τ_j. Therefore, the lasso prior is more flexible than the ridge prior which only relies on the overall shrinkage parameter in (4.8). Figure 4.1 clearly shows that the lasso prior has a sharper peak around zero compared to the ridge prior.

Disadvantages of the lasso

The popularity of the classical lasso lies in its ability to shrink coefficients to zero, thereby automatically performing variable selection. However, there are several disadvantages to the classical lasso. Specifically, (i) it cannot select more predictors than observations, which is problematic when $p > n$; (ii) when a group of predictors is correlated, the lasso generally selects only one predictor of that group; (iii) the prediction error is higher for the lasso compared to the ridge when $n > p$ and the predictors are highly correlated; (iv) it can lead to overshrinkage of large coefficients (see e.g., Polson & Scott, 2011); and (v) it does not always have the oracle property, which implies it does not always perform as well in terms of variable selection as if the true underlying model has been given (Fan & Li, 2001). The lasso only enjoys the oracle property under specific and stringent conditions (Fan & Li, 2001; Zou, 2006). These disadvantages have sparked the development of several generalizations of the lasso. We will now discuss the Bayesian counterparts of several of these generalizations, including the elastic net, group lasso, and hyperlasso. Note that for the Bayesian lasso, coefficients cannot become exactly zero and thus a criterion is needed to select the relevant variables. Depending on the criterion used, more predictors than observations could be selected. However, the Bayesian lasso does not allow a grouping structure to be included, it overshrinks large coefficients, and it does not have the oracle property since the tails for the prior on β_j are not heavier than exponential tails (Polson, Scott, & Windle, 2014).

4.3.1 Elastic net

The most popular generalization of the lasso is the elastic net (Zou & Hastie, 2005). The elastic net can be seen as a combination of the ridge and lasso. The elastic net resolves issues (i), (ii), and (iii) of the ordinary lasso. The elastic net prior can be obtained as the following scale mixture of normals (Li & Lin, 2010):

$$\begin{aligned}
\beta_j|\lambda_2,\tau_j,\sigma^2 &\sim \text{Normal}\left(0,(\frac{\lambda_2}{\sigma^2}\frac{\tau_j}{\tau_j-1})^{-1}\right) \\
\tau_j|\lambda_2,\lambda_1,\sigma^2 &\sim \text{Truncated-Gamma}\left(\frac{1}{2},\frac{8\lambda_2\sigma^2}{\lambda_1^2}\right), \text{ for } j=1,\ldots,p, \\
\lambda_1 &\sim \text{half-Cauchy}(0,1) \\
\lambda_2 &\sim \text{half-Cauchy}(0,1)
\end{aligned} \tag{4.13}$$

where the truncated Gamma density has support $(1,\infty)$. This implies the following conditional prior distributions for the regression coefficients:

$$\begin{aligned}
p(\beta_j|\sigma^2,\lambda_1,\lambda_2) = C(\lambda_1,\lambda_2,\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}(\lambda_1|\beta_j|+\lambda_2\beta_j^2)\right\}, \\
\text{for } j=1,\ldots,p,
\end{aligned} \tag{4.14}$$

where $C(\lambda_1,\lambda_2,\sigma^2)$ denotes the normalizing constant. The corresponding posterior modes for β_j are equivalent to the estimates from the classical elastic net penalty. Expression (4.14) illustrates how the elastic net prior offers a combination of the double-exponential prior, i.e., the lasso penalty $\lambda|\beta_j|$, and the normal prior, i.e., the ridge penalty $\lambda\beta_j^2$. Specifically, the two penalty parameters λ_1 and λ_2 determine the relative influence of the lasso and ridge penalty, respectively. This can also be seen in Figure 4.1: the elastic net is not as sharply peaked as the lasso prior, but it is sharper than the ridge prior. As mentioned in the Introduction, a disadvantage of the classical elastic net is that the sequential cross-validation procedure used to determine the penalty parameters results in overshrinkage of the coefficients. This problem is resolved in the Bayesian approach by estimating both penalty parameters simultaneously through a full or empirical Bayes approach.

4.3.2 Group lasso

The group lasso (M. Yuan & Lin, 2006) is a generalization of the lasso primarily aimed at improving performance when predictors are grouped in some way, for example when qualitative predictors are coded as dummy or one-hot variables (as is often implicitly done in ANOVA, for instance). Similarly to the elastic net, the

penalty function induced by the group lasso lies between the l_1 penalty of the lasso in (4.12) and the l_2 penalty of the ridge in (4.8). To apply the group lasso, the vector of regression coefficients $\boldsymbol{\beta}$ is split in G vectors $\boldsymbol{\beta}_g$, where each vector represents the coefficients of predictors in that group. Denote by m_g the dimension of each vector $\boldsymbol{\beta}_g$. The group lasso corresponds to the following scale mixture of normals (Kyung et al., 2010):

$$\begin{aligned} \boldsymbol{\beta}_g|\tau_g^2, \sigma^2 &\sim \text{MVN}(\mathbf{0}, \sigma^2\tau_g^2 I_{m_g}) \\ \tau_g^2|\lambda^2 &\sim \text{Gamma}(\frac{m_g+1}{2}, \frac{\lambda^2}{2}), \text{ for } g = 1, \ldots, G, \\ \lambda &\sim \text{half-Cauchy}(0, 1) \end{aligned} \tag{4.15}$$

where MVN denotes the multivariate normal distribution with dimension m_g and I_{m_g} denotes an $(m_g \times m_g)$ identity matrix. Note that, contrary to the priors considered thus far, the group lasso prior does not consist of independent priors on the regression coefficients β_j, but rather independent priors on the groups of regression coefficients $\boldsymbol{\beta}_g$. If there is no grouping structure, $m_g = 1$ and the Bayesian group lasso in (4.15) reduces to the Bayesian lasso in (4.11). The scale mixture of normals in (4.15) leads to the following conditional prior for the regression coefficients (Kyung et al., 2010):

$$p(\beta_j|\sigma^2, \lambda) = C \exp\left\{-\frac{\lambda}{\sqrt{\sigma^2}} \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||\right\}, \text{ for } g = 1, \ldots, G, \text{ and } j = 1, \ldots, p, \tag{4.16}$$

where $||\boldsymbol{\beta}_g|| = (\boldsymbol{\beta}_g'\boldsymbol{\beta}_g)^{\frac{1}{2}}$ and C denotes the normalizing constant. Due to the simultaneous penalization of all coefficients in one group, all estimated regression coefficients in one group will be either zero or nonzero, depending on the value for λ.

4.3.3 Hyperlasso

Zou (2006) proposes the adaptive lasso as a generalization of the lasso that enjoys the oracle property (limitation (v) of the lasso), i.e., it performs as well as if the true underlying model has been given. The central idea of the adaptive lasso is to separately weigh the penalty for each coefficient based on the observed data. A Bayesian adaptive lasso has been proposed, among others, by Alhamzawi et al. (2012) and Feng, Wu, and Song (2015). However, as noted by Griffin and Brown (2011), the weights included in the adaptive lasso place great demands on the data,

which can lead to poor performance in terms of prediction and variable selection when the sample size is small. Therefore, Griffin and Brown (2011) propose the hyperlasso as a Bayesian alternative to the adaptive lasso, which is obtained through the following mixture of normals:

$$\begin{aligned} \beta_j|\phi_j^2 &\sim \text{Normal}(0, \phi_j^2) \\ \phi_j^2|\tau_j &\sim \text{Exponential}(\tau_j) \\ \tau_j|\nu, \lambda^2 &\sim \text{Gamma}(\nu, \frac{1}{\lambda^2}) \text{ for } j = 1, \ldots, p. \\ \lambda &\sim \text{half-Cauchy}(0, 1) \end{aligned} \tag{4.17}$$

This is equivalent to placing a Gamma mixing density on the hyperparameter of the double-exponential prior:

$$\begin{aligned} \beta_j|\tau_j &\sim \text{Double-Exponential}(0, (2\tau_j)^{1/2}) \\ \tau_j|\nu, \lambda^2 &\sim \text{Gamma}(\nu, \frac{1}{\lambda^2}), \text{ for } j = 1, \ldots, p. \end{aligned} \tag{4.18}$$

Note that the density of the hyperlasso prior strongly resembles the density of the lasso prior (Figure 4.1), the main difference being that the hyperlasso has heavier tails than the lasso. Contrary to the priors considered thus far, this prior corresponds to a penalty that is non-convex implying that multiple posterior modes can exist. Therefore, care must be taken to ensure that the complete posterior distribution is explored. In addition, the hyperlasso prior for β is not conditioned on the error variance σ^2. Following Griffin and Brown (2011), we will consider the specific case of $\nu = 0.5$. However, whereas Griffin and Brown (2011) use cross-validation to choose λ, we will rely on a full and empirical Bayes approach.

4.3.4 Horseshoe

A popular shrinkage prior in the Bayesian literature is the horseshoe prior (Carvalho et al., 2010):

$$\begin{aligned} \beta_j|\tau_j^2 &\sim \text{Normal}(0, \tau_j^2) \\ \tau_j|\lambda &\sim \text{Half-Cauchy}(0, \lambda), \text{ for } j = 1, \ldots, p \\ \lambda|\sigma &\sim \text{Half-Cauchy}(0, \sigma). \end{aligned} \tag{4.19}$$

Note that Carvalho et al. (2010) explicitly include the half-Cauchy prior for λ in their specification, thereby implying a full Bayes approach. This formulation results

in a horseshoe prior that is automatically scaled by the error standard deviation σ. The half-Cauchy prior can be written as a mixture of inverse Gamma and Gamma densities, so that the horseshoe prior in (4.19) can be equivalently specified as:

$$\begin{aligned} \beta_j|\tau_j^2 &\sim \text{Normal}(0, \tau_j^2) \\ \tau_j^2|\omega &\sim \text{inverse Gamma}(\frac{1}{2}, \omega) \\ \omega|\lambda^2 &\sim \text{Gamma}(\frac{1}{2}, \lambda^2) \\ \lambda^2|\gamma &\sim \text{inverse Gamma}(\frac{1}{2}, \gamma) \\ \gamma|\sigma^2 &\sim \text{Gamma}(\frac{1}{2}, \sigma^2) \end{aligned} \tag{4.20}$$

An expression for the marginal prior of the regression coefficients β_j is not analytically tractable. The name "horseshoe" prior arises from the fact that for fixed values $\lambda = \sigma = 1$, the implied prior for the shrinkage coefficient $\kappa_j = \frac{1}{1+\tau_j^2}$ is similar to a horseshoe shaped Beta(0.5, 0.5) prior. Large coefficients will lead to a shrinkage coefficient κ_j that is close to zero such that there is practically no shrinkage, whereas small coefficients will have a κ_j close to 1 and will be shrunken heavily. Note that the horseshoe prior is the only prior with an asymptote at zero (Figure 4.1). Combined with the heavy tails, this ensures that small coefficients are heavily shrunken towards zero while large coefficients remain large. The horseshoe prior has also been termed a global-local shrinkage prior (e.g., Polson & Scott, 2011) because it has a predictor-specific local shrinkage component τ_j as well as a global shrinkage component λ. The basic intuition is that the global shrinkage parameter λ performs shrinkage on all coefficients and the local shrinkage parameters τ_j loosen the amount of shrinkage for truly large coefficients. Many global-local shrinkage priors (including the horseshoe and hyperlasso) are special cases of the general class of hypergeometric inverted-beta distributions (Polson & Scott, 2012). In addition to the full Bayes approach implied by the specification in (4.19), we will also consider an empirical Bayes approach to determine λ.

4.3.5 Regularized horseshoe

The horseshoe prior in Subsection 4.3.4 has the characteristic that large coefficients will not be shrunken towards zero too heavily. Indeed, this is one of the advertised qualities of the horseshoe prior (Carvalho et al., 2010). Although this property is desirable in theory, it can be problematic in practice, especially when parameters are weakly identified. In this situation, the posterior means of the re-

gression coefficients might not exist and even if they do, the horseshoe prior can result in an unstable MCMC sampler (Ghosh et al., 2018). To solve these problems Piironen and Vehtari (2017b) propose the regularized horseshoe, which is defined as follows:

$$\begin{aligned} \beta_j|\tilde{\tau}_j^2, \lambda &\sim \text{Normal}(0, \tilde{\tau}_j^2\lambda), \text{ with } \tilde{\tau}_j^2 = \frac{c^2\tau_j^2}{c^2 + \lambda^2\tau_j^2} \\ \lambda|\lambda_0^2 &\sim \text{half-Cauchy}(0, \lambda_0^2), \text{ with } \lambda_0 = \frac{p_0}{p - p_0}\frac{\sigma}{\sqrt{n}} \\ \tau_j &\sim \text{half-Cauchy}(0, 1) \\ c^2|\nu, s^2 &\sim \text{inverse Gamma}(\nu/2, \nu s^2/2), \end{aligned} \tag{4.21}$$

where p_0 represents a prior guess of the number of relevant variables. The resulting prior will shrink small coefficients in the same way as the horseshoe[5], but unlike the horseshoe, large coefficients will be shrunken towards zero by a Student's t distribution with ν degrees of freedom and scale s^2. Piironen and Vehtari (2017b) use a Student's t distribution with $\nu = 4$ and $s^2 = 2$ and we use the same hyperparameters, although other choices are possible. As long as the degrees of freedom ν are small enough, the tails will be heavy enough to ensure a robust shrinkage pattern for large coefficients.

It is possible to specify a half-Cauchy prior for the global shrinkage parameter with a scale equal to 1 or the error standard deviation, i.e., $\lambda \sim \text{half-Cauchy}(0, 1)$ or $\lambda \sim \text{half-Cauchy}(0, \sigma)$, for example when no prior information is available regarding the number of relevant variables. However, as noted by Piironen and Vehtari (2017b), the scale based on the a priori number of relevant variables will generally be much smaller than 1 or σ. In addition, even if the prior guess for the number of relevant parameters p_0 is incorrect, the results are robust to this choice as long as a half-Cauchy prior is used. Following the recommendations of Piironen and Vehtari (2017b), we will only consider a full Bayes approach to determine λ in the regularized horseshoe.

4.3.6 Discrete normal mixture

The normal mixture prior is a discrete mixture of a peaked prior around zero (the spike) and a vague proper prior (the slab); it is therefore also termed a spike-and-slab prior. It is substantially different from the priors considered thus far, which are all continuous mixtures of normal densities. Based on the data, regression

[5]Because of the similarity to the horseshoe, the density and contour plots for the regularized horseshoe are not substantially different from those of the horseshoe and are therefore not included in Figure 4.1 and Figure 4.3.

coefficients close to zero will be assigned to the spike, resulting in shrinkage towards 0, while coefficients that deviate substantially from zero will be assigned to the slab, resulting in (almost) no shrinkage. Early proposals of mixture priors can be found in George and McCulloch (1993) and Mitchell and Beauchamp (1988), and a scale mixture of normals formulation can be found in Ishwaran and Rao (2005). We will consider the following specification of the mixture prior:

$$\begin{aligned} \beta_j|\gamma_j, \tau_j^2, \phi_j^2 &\sim (\gamma_j)\text{Normal}(0, \tau_j^2) + (1-\gamma_j)\text{Normal}(0, \phi_j^2) \\ \tau_j^2 &\sim \text{inverse Gamma}(0.5, 0.5), \text{ for } j = 1, \ldots, p, \end{aligned} \tag{4.22}$$

where τ_j is given a vague prior so that the variance of the slab is estimated based on the data and ϕ_j^2 is fixed to a small number, say $\phi_j^2 = 0.001$, to create the spike. By assigning an inverse Gamma(0.5, 0.5) prior on τ_j^2, the resulting marginal distribution of the slab component of the mixture is a Cauchy distribution.

There are several options for the prior on the mixing parameter γ_j. In this paper, we will consider the following two options: 1) γ_j as a Bernoulli distributed variable taking on the value 0 or 1 with probability 0.5, i.e., $\gamma_j \sim$ Bernoulli(0.5); and 2) γ_j uniformly distributed between 0 and 1, i.e., $\gamma_j \sim$ Uniform$(0, 1)$. In the first option, which we label the Bernoulli mixture, each coefficient β_j is given either the slab or the spike as prior. The second option, labelled the uniform mixture, is more flexible in that each coefficient is given a prior consisting of a mixture of the spike and slab, with each component weighted by the uniform probabilities γ_j

The density of the normal mixture prior is presented in Figure 4.1, which clearly shows the prior is a combination of two densities. The representation in Figure 4.1 is based on a normal mixture with equal mixing probabilities, rather than a Bernoulli or uniform prior on the mixing probabilities. Note that the mixture prior is not conditioned on the error variance σ^2. We will only consider a full Bayesian approach for the mixture priors.

4.4 Illustrating the behavior of the shrinkage priors

4.4.1 Contour plots

Contour plots provide an insightful way to illustrate the behavior of classical penalties and Bayesian shrinkage priors. First, consider Figure 4.2 which shows the frequentist and Bayesian contour plots for the lasso. In both plots, the green elliptical lines represent the contours of the sum of squared residuals, centered around the

regular OLS estimate $\hat{\beta}_{OLS}$. The solid black diamond in the left plot represents the constraint region for the classical lasso penalty function for two predictors β_1 and β_2. The classical penalized regression solution $\hat{\beta}_{LASSO}$ is the point where the contour of the sum of squared residuals meets the constraint region. This point corresponds to the minimum of the penalized regression equation in (4.1). In the right plot, the diamond shaped contours reflect the shape of the lasso prior (Section 4.3). The contour of the Bayesian posterior distribution based on the lasso prior is shown in blue. As can be seen, the posterior distribution is located between the sum of squared residuals contour and the prior contour. The Bayesian posterior median estimate $\hat{\beta}_{BAYES}$ is added in blue and shrunken towards zero compared to the OLS estimate $\hat{\beta}_{OLS}$. Note that the posterior mode would correspond to the classical penalized regression solution, if the same value for the penalty parameter λ is used.

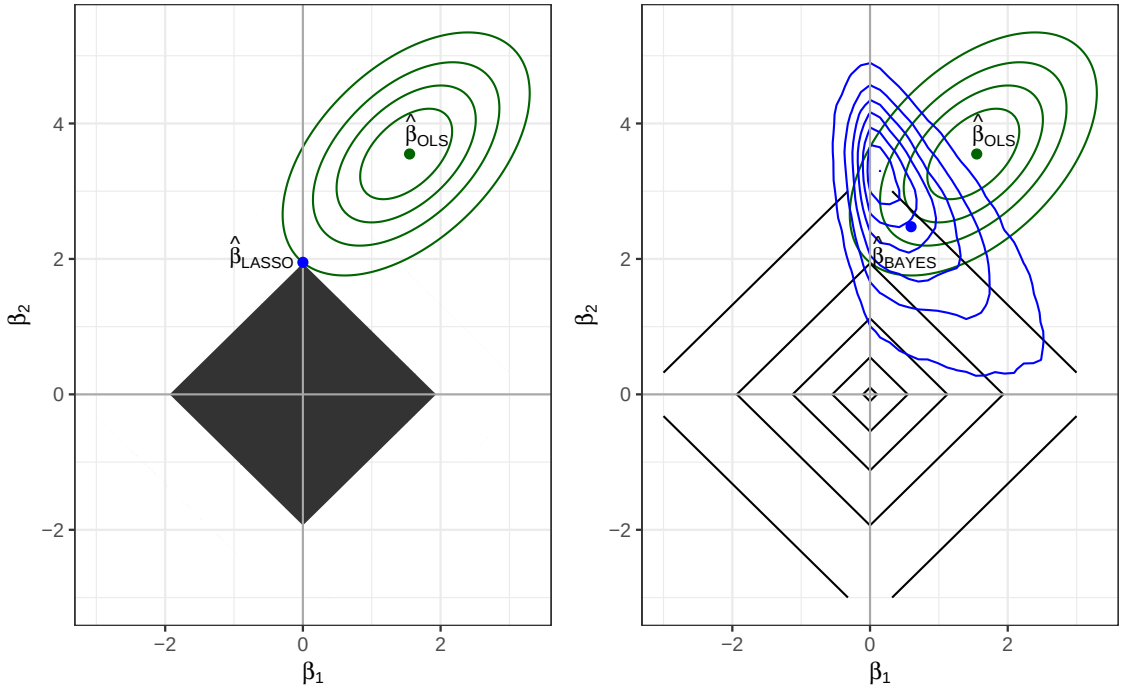

Figure 4.2: Contour plot representing the sum of squared residuals, classical lasso constraint region (left), bivariate lasso prior and posterior distribution (right), and the classical and Bayesian penalized point estimates.

Figure 4.3 shows the contour plots of the different shrinkage priors for two predictors β_1 and β_2, while Figure 4.4 shows the contour plots for the lasso and group lasso for three predictors. From a classical penalization perspective, the lasso and elastic net penalties have sharp corners at $\beta_1 = \beta_2 = 0$. As a result, the contour of the sum of squared residuals will meet the contours of these penalties more easily at a point where one of the coefficients equals zero, which explains why these penalties can shrink coefficients to exactly zero. The ridge penalty, on the other hand, does not show these sharp corners and can therefore not shrink coefficients to exactly

zero. From a Bayesian penalization perspective, the bivariate prior contour plots illustrate the shrinkage behavior of the priors. For example, the hyperlasso and horseshoe have a lot of prior mass where at least one element is close to zero, while the ridge has most prior mass where both elements are close to zero. Figure 4.3 also shows that the ridge, local Student's t, lasso, and elastic net are convex. This can be seen when drawing a straight line from one point to another point on a contour. For a convex distribution, the line lies completely within the contour. The hyperlasso and horseshoe prior are non-convex, which can be seen from the starlike shape of the contour. Frequentist penalization has generally focused on convex penalties, due to their computational convenience for optimization procedures. In the Bayesian framework, which relies on sampling (MCMC) techniques, the use of convex and non-convex priors is computationally similar. It is recommendable, however, to use multiple starting values in the case of non-convex priors due to possible multimodality of the posterior distribution (Griffin & Brown, 2011).

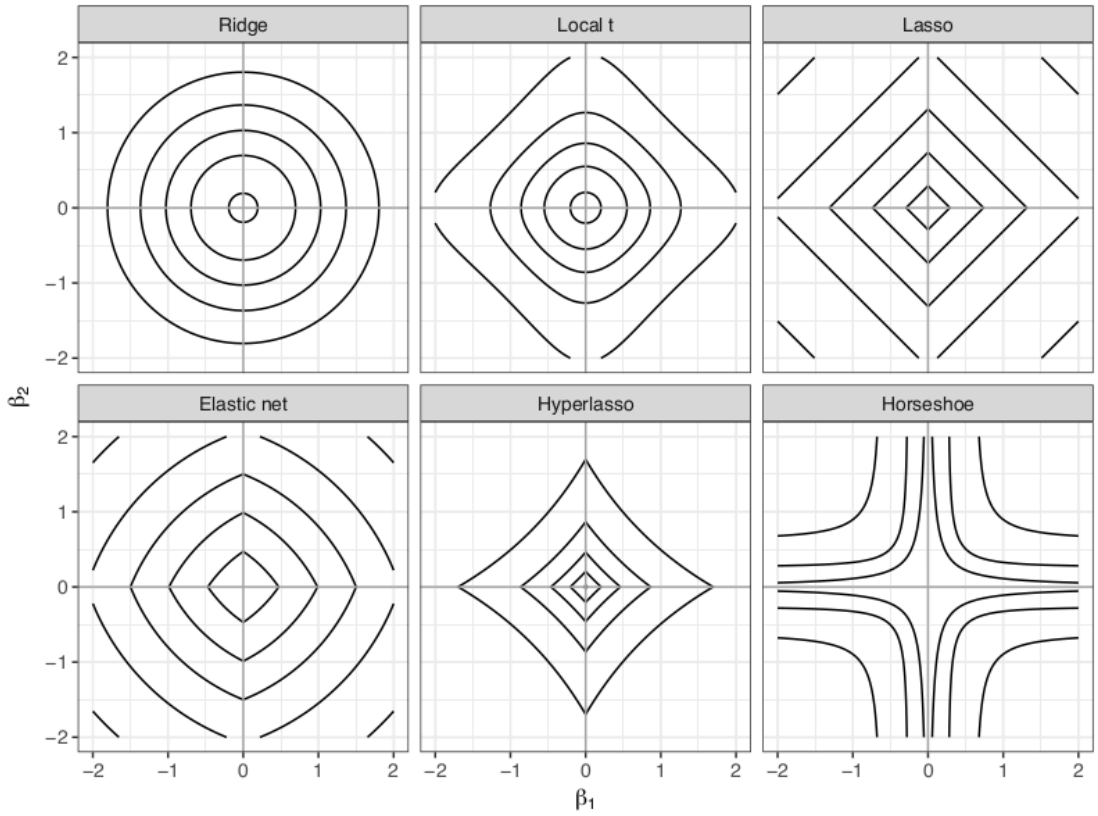

Figure 4.3: Contour plots representing the bivariate prior distribution of the shrinkage priors

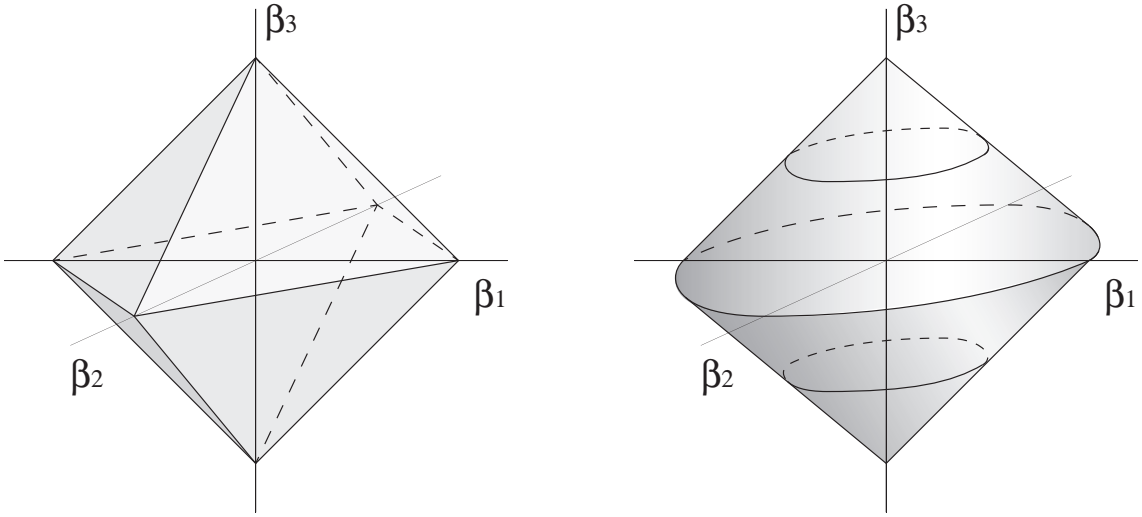

Figure 4.4: Contour plots of the lasso (left) and group lasso (right) in $\mathbf{R}^3$, with β_1 and β_2 belonging to group 1 and β_3 belonging to group 2. For the group lasso, if we consider only β_1 and β_2, which belong to the same group, the contour resembles that of the ridge with most prior mass if both β_1 and β_2 are close to zero. On the other hand, if we consider β_1 and β_3, which belong to different groups, the contour is similar to that of the lasso, which has more prior mass where only one element is close to zero. This illustrates how the group lasso simultaneously shrinks elements belonging to the same group.

4.4.2 Shrinkage behavior

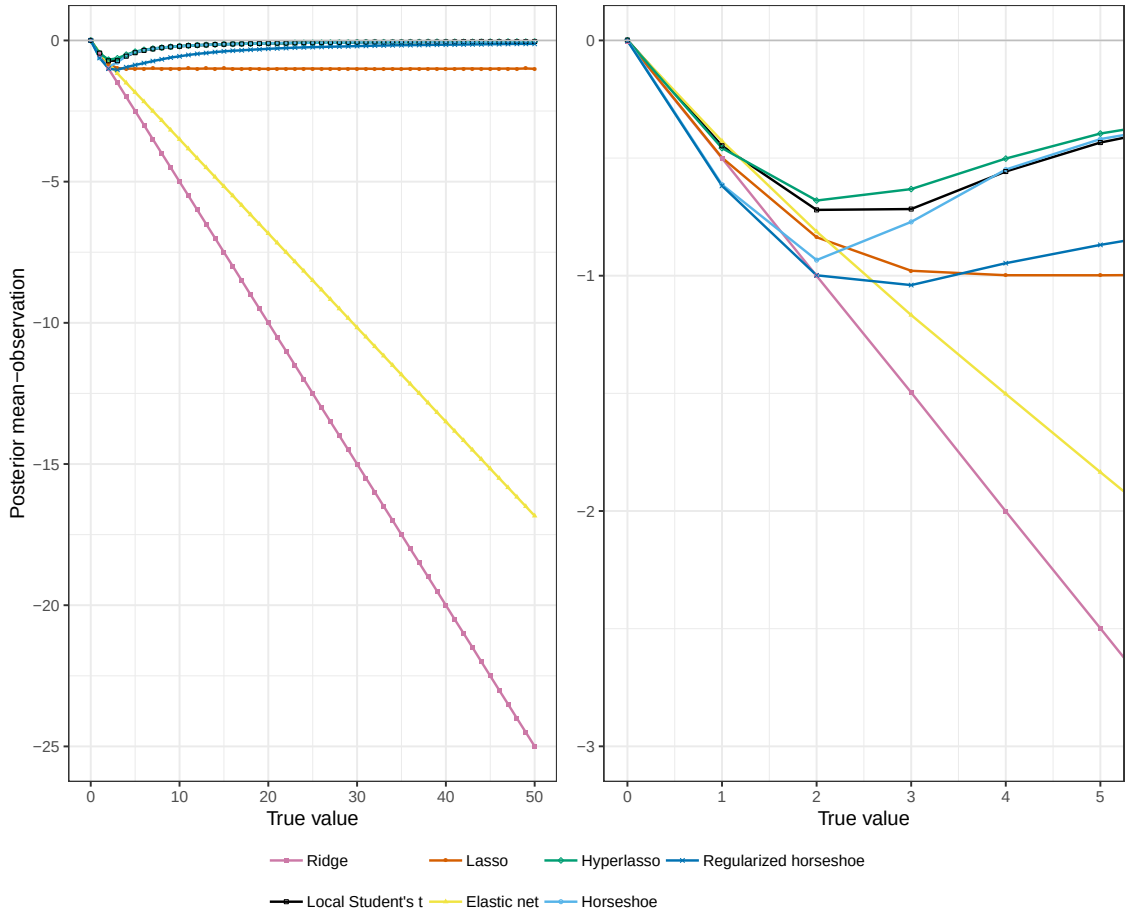

Figure 4.5: Difference between the estimated and true effect for the shrinkage priors in a simple normal model with the penalty parameter λ fixed to 1.

Prior shrinkage of small effects towards zero is important to obtain sparse solutions. Figure 4.5 illustrates the shrinkage behavior of the priors in a simple normal model: $y \sim \text{Normal}(\beta, 1)$. We estimate β based on a single observation y, which is varied from 0 to 50. Using only a single observation is possible because the variance is known. The penalty parameter λ for each shrinkage prior is fixed to 1. The resulting difference between the posterior mean estimates and true means is shown in Figure 4.5. The behavior of the priors varies greatly. Specifically, for the ridge and elastic net priors, the difference between the estimated and true effect increases as the true mean increases. For the lasso prior, the difference increases for small effects and then remains constant. Note how the difference for the elastic net lies between the difference obtained under the ridge and lasso priors, illustrating that the elastic net is a combination of the ridge and lasso priors. The other shrinkage priors all show some differences between estimated and true means for small effects, indicating shrinkage of these effects towards zero, but the difference is practically

zero for large effects. The right column of Figure 4.5 provides the same figure, but zoomed in on the small effects. Note how the regularized horseshoe shrinks large effects more than the horseshoe prior, but goes to zero eventually.

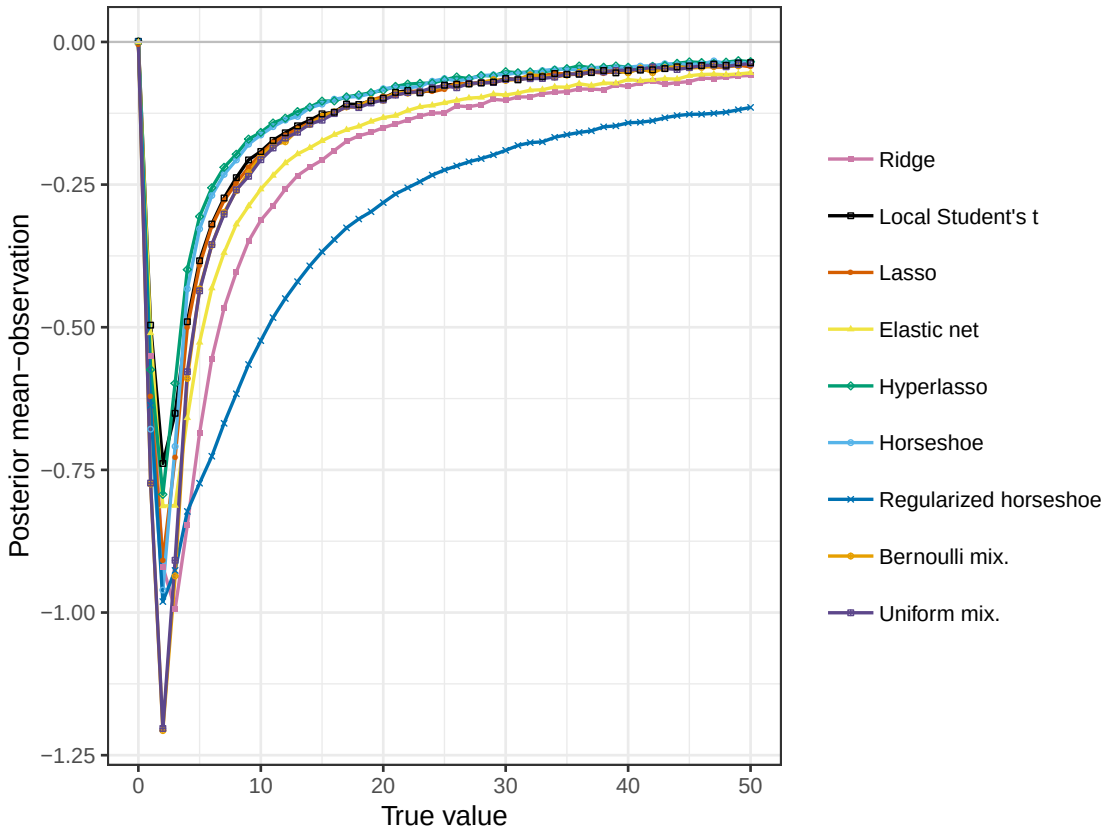

Figure 4.6: Difference between the estimated and true effect for the shrinkage priors in a simple normal model with a half-Cauchy hyperprior specified for the penalty parameter λ.

A similar illustration is presented in Figure 4.6, but based on a full Bayes approach where λ is freely estimated. Thus, instead of fixing λ to a specific value, it is given a standard half-Cauchy prior distribution and estimated simultaneously with the other parameters in the model. Overall, all shrinkage priors show differences between true and estimated means for small effects, which decrease towards zero as the effect grows. Note that the difference is negative, indicating that the estimated mean is smaller than the true mean. Thus, all shrinkage priors heavily pull small effects towards zero, while asserting almost no influence on larger effects, although some shrinkage still occurs even when the true mean equals 50. The mixture priors result in the largest differences between true and estimated small effects, indicating the most shrinkage, and the local Student's t prior shows the smallest difference for small effects. As the effect grows, the regularized horseshoe prior results in estimates farthest from the true effects, indicating the most shrinkage for large effects.

These illustrations indicate that when the penalty parameter is fixed, only the local Student's t, hyperlasso, and (regularized) horseshoe priors allow for shrinkage of small effects while estimating large effects correctly. However, if a prior is specified for the penalty parameter, so that the uncertainty in this parameter is taken into account, all shrinkage priors show this desirable behavior.

4.5 Simulation study

4.5.1 Conditions

We conduct a Monte Carlo simulation study to compare the performance of the shrinkage priors and several frequentist penalization methods. We simulate data from the linear regression model, given by: $\boldsymbol{y} = \beta_0\mathbf{1}+\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, with $\epsilon_i \sim \text{Normal}(0, \sigma^2)$. We consider six simulation conditions. Conditions (1)-(5) are equal to the conditions considered in Li and Lin (2010). In addition, condition (1) and (2) have also been considered in Kyung et al. (2010); Roy and Chakraborty (2016); Tibshirani (1996); Zou and Hastie (2005). Condition (6) has been included to investigate a setting in which $p > N$. The conditions are as follows[6]:

1. $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$; $\sigma^2 = 9$; $\boldsymbol{X}$ generated from a multivariate normal distribution with mean vector $\mathbf{0}$, variances equal to 1, and pairwise correlations between predictors equal to 0.5. The number of observations is $n = 240$, with 40 observations for training and 200 observations for testing the model.

2. $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)'$; the other settings are equal to those in condition (1).

3. $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{15})$; $\sigma^2 = 225$; $\boldsymbol{x}_j = Z_1 + \omega_j$, for $j = 1, \ldots, 5$; $\boldsymbol{x}_j = Z_2 + \omega_j$, for $j = 6, \ldots, 10$; $\boldsymbol{x}_j = Z_3 + \omega_j$, for $j = 11, \ldots, 15$; and $\boldsymbol{x}_j \sim \text{Normal}(0, 1)$, for $j = 16, \ldots, 30$. Here, Z_1, Z_2, and Z_3 are independent standard normal variables and $\omega_j \sim \text{Normal}(0, 0.01)$. The number of observations is $n = 600$, with 200 observations for training and 400 observations for testing the model.

4. The number of observations is $n = 800$, with 400 observations for training and 400 observations for testing the model; the other settings are equal to those in condition (3).

[6]We have also considered two additional conditions in which $p > n$ and the predictors are not highly correlated. Unfortunately, most shrinkage priors resulted in too much non-convergence to trust the results. A description of these additional conditions and the available results for the priors that did obtain enough convergence is available at https://osf.io/nveh3/. Additionally, we would like to refer to Kaseva (2018) where a more sparse, modified version of condition 1 is considered.

5. $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{3, \ldots, 3}_{10})$; the number of observations is $n = 440$, with 40 observations for training and 400 observations for testing the model; the other settings are equal to those in condition (3).

6. The number of observations is $n = 55$, with 25 observations for training and 30 observations for testing the model; the other settings are equal to those in condition (5).

We simulate 500 data sets per condition. All Bayesian methods have been implemented in the software package Stan (Stan Development Team, 2017b), which we call from R using `Rstan` (Stan Development Team, 2018). We include the classical penalization methods available in the R-packages `glmnet` (Friedman, Hastie, & Tibshirani, 2010) and `grpreg` (Breheny & Huang, 2015), i.e., the ridge, lasso, elastic net, and group lasso, for comparison. For the classical penalization methods, the penalty parameter λ is selected based on cross-validation using 10 folds. We also include classical forward selection from the `leaps` (Lumley, 2017) package and we select the model based on three different criteria: the adjusted R^2, Mallows' C_p, and the BIC. For both the Bayesian and the classical group lasso, a grouping structure should be supplied for the analysis. We have used the grouping structure under which the data was simulated. Thus, for conditions 3 until 6, we have four groups with the following regression coefficient belonging to each group: $G_1 = \beta_1, \ldots, \beta_5$, $G_2 = \beta_6, \ldots, \beta_{10}$, $G_3 = \beta_{11}, \ldots, \beta_{15}$, and $G_4 = \beta_{16}, \ldots, \beta_{30}$. All code for the simulation study is available at `https://osf.io/bf5up/`.

4.5.2 Outcomes

The two main goals of regression analysis are: (1) to select variables that are relevant for predicting the outcome, and (2) to accurately predict the outcome. Therefore, we will focus on the performance of the shrinkage priors in terms of variable selection and prediction accuracy. Unlike frequentist penalization methods, Bayesian penalization methods do not automatically shrink regression coefficients to be exactly zero. A criterion is thus needed to select the relevant variables, for which we will use the credibility interval criterion.[7] Using the credibility interval criterion, a predictor is excluded when the credibility interval for β_j covers 0, and it is included when 0 is not contained in the credibility interval. This criterion thus

[7]We have also considered the scaled neighborhood criterion (Li & Lin, 2010) and a fixed cut-off value to select the predictors. The scaled neighborhood criterion excludes a predictor if the posterior probability contained in $[-\sqrt{\text{var}(\beta_p|\mathbf{y})}, \sqrt{\text{var}(\beta_p|\mathbf{y})}]$ exceeds a certain threshold. However, this criterion generally performed worse than the credibility interval criterion. For the fixed cut-off value we excluded predictors when the posterior estimate $|\hat{\beta}| \leq 0.1$ based on Feng et al. (2015). However, the choice of this threshold is rather arbitrary and resulted in very high false inclusion rates.

depends on the percentage of posterior probability mass included in the credibility interval. We will investigate credibility intervals ranging from 0 to 100%, with steps of 10%. The optimal credibility interval is selected using the distance criterion (see e.g., Perkins & Schisterman, 2006), i.e.,

$$\text{distance} = \sqrt{(1 - \text{correct inclusion rate})^2 + (\text{false inclusion rate})^2}, \tag{4.23}$$

The credibility interval with the lowest distance is optimal in terms of the highest correct inclusion rate and lowest false inclusion rate. For the selected credibility interval, we will report Matthews' correlation coefficient (MCC; Matthews, 1975), which is a measure indicating the quality of the classification. MCC ranges between -1 and +1 with MCC = -1 indicating complete disagreement between the observed and predicted classifications and MCC = +1 indicating complete agreement.

To assess the prediction accuracy of the shrinkage priors, we will consider the prediction mean squared error (PMSE) for each replication. To compute the PMSE, we first estimate the regression coefficients $\hat{\boldsymbol{\beta}}$ on the training data only. These estimates are then used to predict the responses on the outcome variable of the test set, $\boldsymbol{y}^{gen}$, for which the actual responses, $\boldsymbol{y}$, are available. Prediction of $\boldsymbol{y}^{gen}$ occurs within the "generated quantities" block in Stan, meaning that for each MCMC draw, y_i^{gen} is generated such that we obtain the full posterior distribution for each y_i^{gen}. The mean of this posterior distribution is used as estimate for y_i^{gen}. The PMSE for each replication can then be computed as: $\frac{1}{N}\Sigma_{i=1}^{N}(y_i^{gen} - y_i)^2$. For each condition, this will result in 500 PMSEs, one for each replication, of which we will compute the median. Furthermore, to assess the uncertainty in the median PMSE estimate, we will bootstrap the standard error (SE) by resampling 500 PMSEs from the obtained PMSE values and computing the median. This process is repeated 500 times and the standard deviation of the 500 bootstrapped median PMSEs is used as SE of the median PMSE.

4.5.3 Convergence

Convergence will be assessed using split $\hat{R}$, which is a version of the often used potential scale reduction factor (PSRF; Gelman & Rubin, 1992) that is implemented in Stan (Stan Development Team, 2017a, p. 370-373). Additionally, Stan reports the number of divergent transitions. A divergent transition indicates that the approximation error in the algorithm accumulates (Betancourt, 2017; Monnahan, Thorson, & Branch, 2016), which can be caused by a too large step size, or because of strong curvature in the posterior distribution. As a result, it can be necessary to adapt the settings of the algorithm or to reparametrize the model. For the simulation, we initially employed a very small step size (0.001) and high tar-

get acceptance rate (0.999), however, these settings result in much slower sampling. Therefore, in the later conditions we used the default step size (1) and a lower target acceptance rate (0.85) and only reran the replications that did not converge with the stricter settings (i.e., smaller step size and higher target acceptance rate). Only if all parameters had a PSRF < 1.1 and there were no divergent transitions, did we consider a replication as converged.[8] We have only included those conditions in the results with at least 50% convergence (i.e., at least 250 converged replications). The convergence rates are available at https://osf.io/nveh3/.

[8]For the horseshoe prior, all replications in all conditions resulted in one or more divergent transitions, despite reparametrization of the model. The regularized horseshoe also resulted in divergent transitions for most replications, although the percentage of divergent transitions was on average much lower for the regularized horseshoe compared to the horseshoe. The percentages divergent transitions are available at https://osf.io/nveh3/. To be able to include these priors in the overview, we have only considered the PSRF to assess convergence and manually checked the traceplots. However, see Kaseva (2018) for a deeper investigation into the divergent transitions and alternative parametrizations of the horseshoe prior.

4.5.4 Prediction accuracy

Table 4.2: Median prediction mean squared error (PMSE) with bootstrapped standard errors in brackets for the shrinkage priors.

Prior	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
			Full Bayes			
Ridge	10.95 (0.08)	10.49 (0.09)	243.09 (0.86)	236.07 (0.95)	319.84 (1.6)	371.61 (7.14)
Local Student's t	10.83 (0.09)	10.71 (0.08)	242.79 (0.8)	236.04 (0.89)	317.72 (2.31)	359.32 (6.67)
Lasso	10.78 (0.07)	10.53 (0.09)	238.89 (0.84)	234.29 (0.99)	316.43 (2.14)	360.37 (5.84)
Elastic net	10.93 (0.08)	10.62 (0.08)	243.23 (0.86)	236.1 (0.93)	323.24 (1.86)	387.47 (6.12)
Group lasso	NA[1]	NA[1]	241.06 (0.84)	235.35 (0.87)	316.23 (2.14)	358.17 (7.14)
Hyperlasso	10.77 (0.09)	10.52 (0.08)	238.69 (0.78)	234.24 (0.97)	316.29 (2.31)	356.45 (5.28)
Horseshoe	10.68 (0.07)	10.88 (0.09)	231.97 (0.87)	230.20 (0.82)	316.56 (2.43)	355.69 (4.18)
Regularized horseshoe true p_0[2]	10.69 (0.07)	10.58 (0.08)	233.42 (0.97)	230.43 (0.84)	316.51 (2.03)	356.93 (4.37)
Bernoulli mixture	***10.57*** (0.10)	11.26 (0.09)	***230.34*** (0.78)	***229.12*** (0.94)	322.35 (2.27)	357.62 (4.89)
Uniform mixture	***10.57*** (0.10)	11.24 (0.09)	230.42 (0.75)	229.36 (0.91)	322.40 (2.03)	359.25 (4.76)
			Empirical Bayes			
Ridge	10.96 (0.08)	*10.29* (0.1)	242.71 (0.9)	235.91 (0.92)	317.85 (2.04)	421.86 (11.93)
Local Student's t	10.97 (0.08)	10.30 (0.08)	242.85 (0.89)	236.05 (0.88)	317.33 (2.12)	375.99 (6.85)
Lasso	10.78 (0.09)	10.49 (0.08)	238.86 (0.77)	234.10 (0.95)	317.38 (2.02)	424.82 (12.56)
Elastic net	10.95 (0.09)	10.31 (0.09)	242.67 (0.9)	235.92 (0.88)	316.64 (1.79)	365.52 (5.91)
Group lasso	NA[1]	NA[1]	240.89 (0.78)	235.29 (0.92)	315.05 (2.21)	369.00 (6)
Hyperlasso	10.79 (0.08)	10.43 (0.07)	238.62 (0.8)	234.02 (0.99)	***314.44*** (2.55)	436.24 (12.02)
Horseshoe	10.67 (0.08)	10.87 (0.08)	231.55 (0.89)	229.79 (0.82)	320.73 (3.08)	***354.69*** (4.13)
			Classical penalization			
Ridge	10.96 (0.07)	**10.11** (0.06)	241.64 (1.13)	235.25 (1.04)	318.44 (2.48)	494.50 (7.29)
Lasso	10.70 (0.09)	11.06 (0.07)	235.76 (0.90)	231.23 (1.07)	339.09 (3.44)	410.84 (7.33)
Elastic net	10.72 (0.08)	10.89 (0.07)	235.96 (0.88)	231.54 (0.93)	335.50 (3.28)	394.76 (7.78)
Group lasso	NA[1]	NA[1]	233.11 (0.70)	229.76 (0.98)	343.06 (3.14)	407.14 (7.47)
			Forward selection			
BIC	11.01 (0.07)	13.08 (0.11)	681.09 (3.95)	679.33 (2.85)	379.59 (4.48)	417.35 (8.34)
Mallows' C_p	11.06 (0.10)	12.45 (0.11)	464.94 (3.81)	466.25 (3.09)	478.00 (33.33)	687.46 (12.26)
Adjusted R^2	11.31 (0.12)	11.83 (0.10)	248.51 (2.35)	241.03 (1.98)	357.79 (3.20)	402.58 (8.73)

Note.
[1] No results are available for the group lasso in condition 1 and 2, since no grouping structure is present in these conditions.
[2] p_0 denotes the prior guess for the number of relevant variables, which was set to the true number of relevant variables, except in condition 2 where all eight variables are relevant, so we set $p_0 = 7$.
[3] The smallest median PMSE per condition across methods is shown in bold and the smallest median PMSE per condition for the Bayesian methods is shown in italics.

Table 4.2 shows the median PMSE per condition for the shrinkage priors and classical penalization methods. For the regularized horseshoe, the prior guess for the number of relevant variables p_0 was based on the data-generating model, however, the results were comparable when no prior guess or an incorrect prior guess was used. For all methods, the median PMSE increases as the condition becomes more complex. The smallest median PMSE per condition across methods is shown in bold and the smallest median PMSE per condition for the Bayesian methods is shown in italics. In condition 1, 3, and 4 the full Bayesian Bernoulli mixture prior performs best; in condition 2, the classical ridge performs best; in condition 5, the empirical Bayesian hyperlasso performs best; and in condition 6, the empirical Bayesian horseshoe performs best. However, the differences between the methods

are relatively small. Only in condition 6, where the number of predictors is larger than the number of observations, the differences between the methods in terms of PMSE become more pronounced. As expected, forward selection performs the worst, especially when Mallows' C_p or the BIC is used to select the best model. This illustrates the advantage of using penalization, even when $p < n$. Overall, we can conclude that in terms of prediction accuracy the penalization methods perform quite similarly, except when $p > n$.[9]

4.5.5 Variable selection accuracy

Table 4.3 shows MCC and the correct and false inclusion rates for the optimal CIs for the shrinkage priors and MCC and the inclusion rates for the classical penalization methods, which automatically select predictors. The bold values indicate the best inclusion rates across all methods, whereas the italic values indicate the best inclusion rates across the Bayesian methods. Again, for the regularized horseshoe the results were comparable regardless of whether a correct, incorrect, or no prior guess was used. In the first condition, the classical penalization methods outperform the Bayesian methods in terms of correct inclusion rates, but at the cost of higher false inclusion rates. This is a well known problem of the lasso and elastic net when cross-validation is used to select the penalty parameter λ. A solution to this problem is to use stability selection to determine λ (Meinshausen & Bühlmann, 2010). The optimal Bayesian methods in the first condition based on the highest value for MCC are the mixture priors, both of which have reasonable correct and false inclusion rates. Note that, generally, the differences with the other Bayesian methods are relatively small in condition 1. In condition 3 and 4, the correct inclusion rates are generally high and the false inclusion rates are increased as well. As a result, the optimal Bayesian methods in condition 3 show a trade-off between correct and false inclusion rates, with the empirical Bayes group lasso having the highest value for MCC. However, the differences in MCC between most Bayesian methods are small and MCC is generally lower compared to condition 1 due to the increased false inclusion rates. In condition 4, multiple methods show a correct inclusion rate of 1, combined with a high false inclusion rate. In terms of MCC, the empirical Bayes ridge prior performs best. In condition 5, both rates and thus the MCC values are slightly lower across all methods, which is a result of the optimal CI being smaller. The full Bayes lasso and regularized horseshoe perform best in terms of MCC, although the other shrinkage priors show comparable MCC values. Condition 6 shows the most pronounced differences between the methods and the

[9]We have also computed the PMSE for a large test set with 1,000,000 observations as an approximation to the theoretical prediction error. In general, the theoretical PMSEs did not differ substantially from the PMSE in Table 4.2, except in condition 6 where the theoretical PMSE was generally larger. The theoretical PMSEs are available online at https://osf.io/nveh3/

greatest trade-off between correct and false inclusion rates. None of the Bayesian methods attain a value for the MCC greater than 0.51, and some shrinkage priors (i.e., the empirical Bayes ridge and lasso) result in a MCC value of only 0.28. In conclusion, although there exist differences between the methods in terms of variable selection accuracy, there is not one method that performs substantially better than the other methods in terms of both correct and false inclusion rates.

Table 4.3: Matthews' correlation coefficient (MCC) and correct and false inclusion rates based on the optimal credibility intervals (CIs) selected using the distance criterion.

	Condition 1				Condition 3				Condition 4				Condition 5				Condition 6			
Prior	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion
Full Bayes																				
Ridge	90	0.78	0.852	0.087	60	0.66	0.993	0.385	60	0.67	***1.000***	0.387	40	0.57	0.826	0.259	30	0.50	0.829	0.341
Local Student's t	80	0.76	0.884	0.132	60	0.67	0.998	0.381	70	0.60	0.875	0.291	40	0.58	0.827	0.246	30	***0.51***	0.820	0.318
Lasso	80	0.77	0.887	0.126	50	0.64	0.997	0.421	50	0.63	***1.000***	0.438	30	***0.59***	*0.863*	0.283	30	0.50	0.788	0.284
Elastic net	90	0.77	0.851	0.094	60	0.63	0.973	0.386	60	0.67	0.999	0.387	40	0.55	0.824	0.272	30	0.49	***0.836***	0.361
Group lasso	NA[1]	NA[1]	NA[1]	NA[1]	60	0.67	0.987	0.364	60	0.68	***1.000***	0.375	40	0.57	0.822	0.246	30	***0.51***	0.824	0.320
Hyperlasso	80	0.77	0.880	0.117	50	0.64	*0.999*	0.419	50	0.63	***1.000***	0.434	40	0.56	0.773	0.202	30	0.50	0.764	0.253
Horseshoe	70	0.78	0.886	0.116	20	0.49	*0.999*	0.609	20	0.49	***1.000***	0.603	20	0.56	0.858	0.311	20	0.50	0.795	0.294
Regularized horseshoe true p_0[2]	70	0.78	0.889	0.118	40	0.64	0.949	0.351	30	0.61	***1.000***	0.453	40	***0.59***	0.794	0.193	30	***0.51***	0.771	0.247
Bernoulli mixture	50	***0.80***	*0.893*	0.099	20	0.66	0.992	0.381	20	0.66	0.993	0.388	20	0.55	0.735	*0.159*	20	0.48	0.627	0.127
Uniform mixture	50	***0.80***	0.889	0.100	20	0.66	0.989	0.381	20	0.65	0.990	0.390	20	0.55	0.733	0.160	20	0.48	0.628	*0.123*
Empirical Bayes																				
Ridge	90	0.78	0.847	0.081	70	0.52	0.781	*0.276*	70	*0.72*	0.982	0.289	40	0.57	0.819	0.240	30	0.28	0.690	0.222
Local Student's t	90	0.79	0.845	***0.080***	60	0.67	*0.999*	0.377	70	0.70	0.949	0.291	40	0.58	0.821	0.241	30	0.49	0.764	0.279
Lasso	80	0.77	0.885	0.118	50	0.64	*0.999*	0.417	50	0.63	***1.000***	0.433	30	0.57	0.831	0.270	20	0.28	0.591	0.273
Elastic net	90	0.78	0.848	0.085	60	0.67	*0.999*	0.377	70	0.71	0.968	0.290	40	0.58	0.834	0.251	30	0.50	0.810	0.314
Group lasso	NA[1]	NA[1]	NA[1]	NA[1]	60	*0.68*	0.998	0.361	70	0.60	0.880	**0.278**	40	0.58	0.820	0.240	30	0.46	0.728	0.277
Hyperlasso	80	0.78	0.883	0.109	50	0.65	*0.999*	0.414	50	0.63	***1.000***	0.431	40	0.53	0.767	0.199	20	0.32	0.575	0.268
Horseshoe	70	0.78	0.876	0.108	20	0.51	0.997	0.578	20	0.51	0.997	0.576	20	0.53	0.840	0.308	20	0.48	0.756	0.266
Classical penalization																				
Lasso	NA[3]	0.72	0.923	0.210	NA[3]	0.66	0.642	**0.023**	NA[3]	0.67	0.632	*0.008*	NA[3]	0.33	0.418	**0.108**	NA[3]	0.27	0.368	0.116
Elastic net	NA[3]	0.62	**0.971**	0.361	NA[3]	**0.97**	**1.000**	0.031	NA[3]	**0.99**	**1.000**	0.013	NA[3]	0.47	0.645	0.161	NA[3]	0.39	0.548	0.153
Group lasso	NA[1]	NA[1]	NA[1]	NA[1]	NA[3]	0.52	**1.000**	0.482	NA[3]	0.50	**1.000**	0.500	NA[3]	0.34	**0.893**	0.603	NA[3]	0.37	0.787	0.462
Forward selection																				
BIC	NA[3]	0.77	0.843	0.093	NA[3]	-0.087	0.086	0.139	NA[3]	-0.092	0.081	0.137	NA[3]	-0.056	0.171	0.213	NA[3]	-0.0056	0.252	0.249
Mallows' C_p	NA[3]	0.72	0.895	0.176	NA[3]	0.023	0.188	0.164	NA[3]	0.023	0.186	0.164	NA[3]	-0.071	0.122	0.172	NA[3]	-0.074	0.024	**0.052**
Adjusted R_2	NA[3]	0.60	0.938	0.343	NA[3]	0.14	0.332	0.201	NA[3]	0.15	0.329	0.197	NA[3]	0.0085	0.338	0.328	NA[3]	0.053	0.378	0.323

Note.

[1] No results are available for the group lasso in condition 1, since no grouping structure is present in these conditions.

[2] p_0 denotes the prior guess for the number of relevant variables, which was set to the true number of relevant variables, except in condition 2 where all eight variables are relevant, so we set $p_0 = 7$.

[3] For the classical penalization methods, the lasso, elastic net, and forward selection automatically shrink some coefficients to exactly zero so that no criterion such as a confidence interval for variable selection is needed. The ridge is not included, since it does not automatically shrink coefficients to zero and therefore always has a correct and false inclusion rate of 1.

[4] The highest MCC and correct inclusion rate and the lowest false inclusion rate per condition across methods are shown in bold and the highest MCC and best rates per condition for the Bayesian methods are shown in italics.

4.6 Empirical applications

We will now illustrate the shrinkage priors on two empirical data sets. An R package `bayesreg` is available online (https://github.com/sara-vanerp/bayesreg) that can be used to apply the shrinkage priors. The first illustration (math performance) shows the benefits of using shrinkage priors in a situation where the number of predictors is smaller than the number of observations. In the second illustration (communities and crime), the number of predictors is larger than the number of observations, and it is necessary to use some form of regularization in order to fit the model.

4.6.1 Math performance

In this illustration, we aim to predict the final math grade of 395 Portuegese students in secondary schools (Cortez & Silva, 2008), obtained from the UCL machine learning repository[10] (Lichman, 2013). The data set includes 30 predictors covering demographic, social and school related characteristics, such as parents' education and the time spent studying. The continuous predictors were standardized and dummy variables were used for the categorical predictors, resulting in a total of 39 predictors. We split the data into an approximately equal training ($n = 197$) and test ($n = 198$) set.

Table 4.4 presents the computation time in seconds for each method, the prediction mean squared error, and the number of included predictors. We have not included the results for the horseshoe prior because this prior resulted in divergent transitions, which in turn led to instable results (specifically, the PMSE varied greatly when rerunning the analysis).

[10]The data is available at
https://archive.ics.uci.edu/ml/datasets/Student+Performance

Table 4.4: Computation time in seconds (with a 2.8 GHz Intel Core i7 processor), prediction mean squared error (PMSE), and number of included predictors for the different methods for the math performance application

Shrinkage prior	Computation time (seconds)	PMSE	Number of included predictors
	Full Bayes		
Ridge	179	19.53	22
Local Student's t	361	19.44	22
Lasso	219	19.25	22
Elastic net	354	19.53	23
Group lasso	342	19.36	22
Hyperlasso	199	19.18	19
Regularized horseshoe with p_0	1474	19.12	17
Bernoulli mixture	24524	19.31	9
Uniform mixture	4370	19.28	9
	Empirical Bayes		
Ridge	341	19.42	22
Local Student's t	443	19.47	22
Lasso	444	19.26	22
Elastic net	603	19.41	22
Group lasso	534	19.50	22
Hyperlasso	387	19.11	19
	Classical penalization		
Ordinary least squares (OLS)	0.013	22.56	39
Ridge	0.118	18.95	38
Lasso	0.0.072	19.11	20
Elastic net	0.0.053	19.25	20
Group lasso	0.187	19.43	21
	Forward selection		
BIC	0.006	21.13	4
Mallows' C_p	0.006	21.42	13
Adjusted R^2	0.006	22.44	25

It is clear that the Bayesian methods are computationally much more intensive than the classical penalization methods, especially the regularized horseshoe and mixture priors. The advantage brought by this increased computation time, however, is the more straightforward interpretation of results such as credibility intervals and the automatic computation of uncertainty estimates. This can be seen in Figure 4.7 which shows the posterior density for one regression coefficient β_1 using the lasso prior and its 95% credibility interval (i.e., the shaded dark blue area). The bootstrapped 95% confidence interval obtained using the `HDCI` package (Liu, Xu, & Li, 2017) in R is shown by the dashed grey lines and can be seen to underestimate the uncertainty. This problem is often observed using classical lasso estimation (Kyung et al., 2010). The PMSE clearly illustrates the advantage of penalization,

even though the number of predictors is not greater than the sample size. Compared to regression using OLS, all penalization methods show lower PMSEs. Moreover, all penalization methods outperform forward selection in terms of PMSE. Between the different penalization methods, differences in PMSE are small.

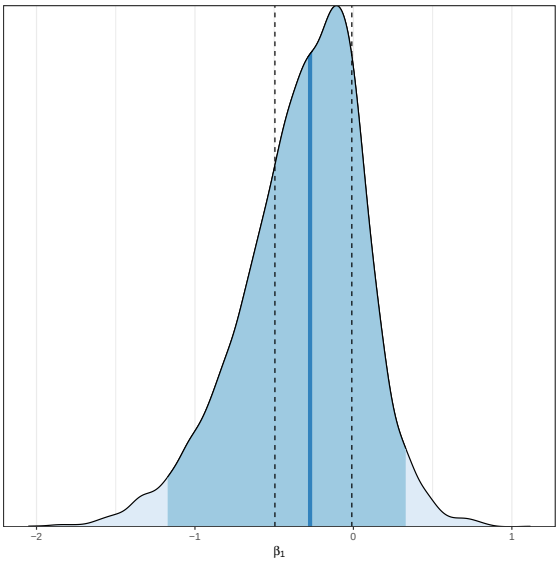

Figure 4.7: Posterior density for β_1 in the math performance application using the Bayesian lasso. The dark blue line depicts the posterior median, the shaded dark blue area depicts the 95% credibility interval. The black dashed lines depict the bootstrapped 95% confidence interval of the classical lasso.

The last column in Table 4.4 reports the number of included predictors for each method. Figure 4.8 shows which predictors are included for each method. Each point indicates an included predictor, based on the optimal CI from condition 5 in the simulation study. OLS does not exclude any predictors, neither does the classical ridge generally although in this data set one coefficient was estimated to be zero. Most shrinkage priors included 22 predictors, with the hyperlasso and regularized horseshoe resulting in a slightly sparser solution. The mixture priors selected much less predictors (9) compared to the other methods. The number of included predictors for the forward selection method ranged from 4 to 25, depending on the criterium used to select the best model.

Based on the predicted errors and the number of included predictors, we conclude that essentially all Bayesian methods and the classical penalization methods performed best. The computation time for the Bayesian methods was considerably larger than for the classical methods. However, this increased computation time results in automatic availability of uncertainty estimates which were generally larger compared to classical bootstrapped confidence intervals.

4.6.2 Communities and crime

We illustrate the shrinkage priors on a data set containing 125 predictors of the number of violent crimes per 100,000 residents in different communities in the US (Redmond & Baveja, 2002) obtained from the UCL machine learning repository[11] (Lichman, 2013). The predictor variables include community characteristics, such as the median family income and the percentage of housing that is occupied, as well as law enforcement characteristics, such as the number of police officers and the police operating budget. We created dummy variables for the two nominal predictors in the data set, resulting in a total of 172 predictors. For the group lasso, all dummy variables corresponding to one predictor make up a group. The number of observations is 319, after removing all cases with at least one missing value on any of the predictors.[12] We split the data into approximately equal training ($n = 159$) and test ($n = 160$) sets. All predictors were normalized to have zero mean and unit variance and the outcome variable was log transformed.

Table 4.5 reports the computation time in seconds for each method, as well as the PMSE and the number of selected variables. Again, the horseshoe prior resulted in divergent transitions and is therefore excluded from the results. The posterior density using the lasso prior for β_{15} is shown in Figure 4.9, with the dark blue shaded

[11] We used the unnormalized data, available at
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized

[12] Although the Bayesian framework allows for straightforward imputation of missing values, we removed all cases with missing values to provide an illustration of the shrinkage methods in a sparse data set.

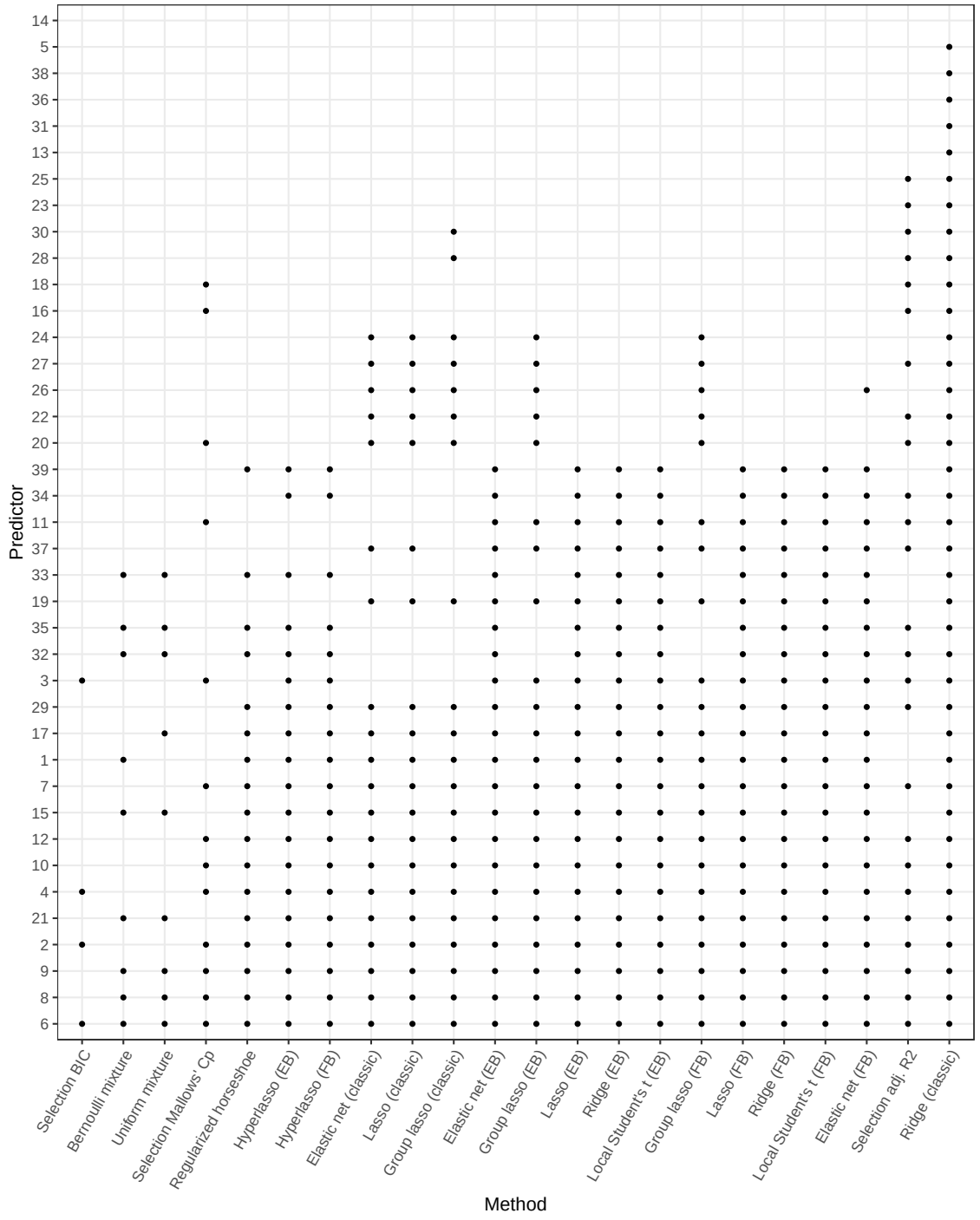

Figure 4.8: Overview of the included predictors for each method in the math performance application. Points indicate that a predictor is included based on the optimal credibility interval (CI) from condition 5 in the simulation study. The methods on the x-axis are ordered such that the method that includes the least predictors is on the left and the method that includes the most predictors is on the right. The predictors on the y-axis are ordered with the predictor being included the least on top and the predictor being included the most at the bottom.

area depicting the 95% credibility intervals and the dashed black lines depicting the bootstrapped 95% confidence interval of the classical lasso. Again, the bootstrapped confidence interval is much smaller than the Bayesian credibility interval and located far from the posterior median estimate (i.e., the dark blue line).

Table 4.5: Computation time in seconds (with a 2.8 GHz Intel Core i7 processor), prediction mean squared error (PMSE), and number of included predictors for the different methods for the crime application

Shrinkage prior	Computation time (seconds)	PMSE	Number of included predictors
	Full Bayes		
Ridge	677	0.217	61
Local Student's t	1973	0.216	60
Lasso	2068	0.216	46
Elastic net	242	0.216	62
Group lasso	3044	0.216	61
Hyperlasso	1066	0.215	46
Regularized horseshoe with p_0	15803	0.226	31
Bernoulli mixture	60006	1.706	54
Uniform mixture	26080	1.683	54
	Empirical Bayes		
Ridge	1195	0.218	60
Local Student's t	1912	0.216	57
Lasso	4207	0.215	46
Elastic net	417	0.217	57
Group lasso	3992	0.217	62
Hyperlasso	2016	0.215	46
	Classical penalization		
Ridge	0.376	0.258	160
Lasso	0.200	0.508	33
Elastic net	0.164	0.460	26
Group lasso	0.408	0.663	55
	Forward selection		
BIC	0.023	1.500	17
Mallows' C_p	0.023	0.276	1
Adjusted R^2	0.023	4.093	141

In addition, most Bayesian methods resulted in a lower PMSE than the classical methods, except for the mixture priors. The forward selection method resulted in much larger PMSEs, except when Mallows' C_p was used to find the best model, however, this model retained only 1 predictor. On the other hand, using the Adjusted R^2 criterion led to a model that included 141 predictors. This illustrates the arbitrariness of using forward selection. Figure 4.10 shows which predictors are included for each method. Each point indicates an included predictor, based on the

optimal CI from condition 6 in the simulation study. Apart from the forward selection methods, the classical elastic net excludes most predictors. Interestingly, the Bayesian elastic net and lasso retain many more predictors than the classical elastic net and lasso. However, not all predictors that are retained by the classical lasso and elastic net are also retained by the Bayesian lasso and elastic net. Specifically, the predictors included by the classical methods but not by the Bayesian methods all correspond to dummy variables for State. The hyperlasso and lasso methods all include 46 predictors, whereas the ridge, local Student's t, elastic net, and group lasso priors all retain around 60 predictors. The mixture priors both include 54 predictors. The regularized horseshoe retains the least predictors of all Bayesian methods, only 31. The classical ridge retains almost all predictors, but estimated some coefficients to be equal to zero in this data set.

Based on this illustration, we conclude that the Bayesian penalization methods outperform the classical penalization methods in terms of prediction error. The prediction errors of the Bayesian penalization methods do not differ substantially, except for the mixture priors which showed larger PMSEs. The shrinkage priors differ in how much shrinkage they perform and thus in the number of predictors that are selected.

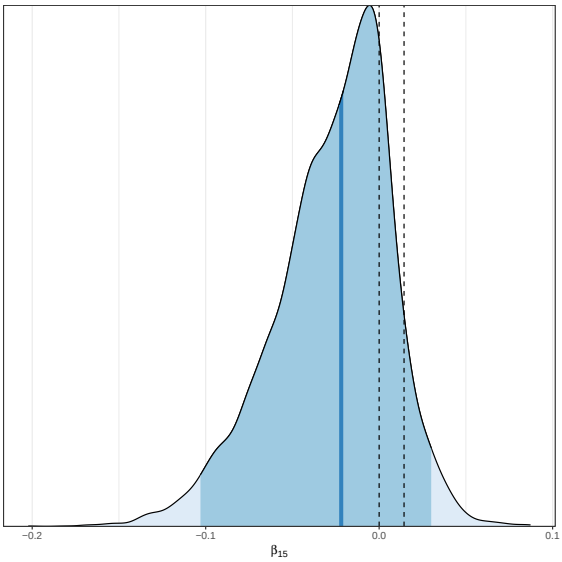

Figure 4.9: Posterior density for β_{15} in the crime application using the Bayesian lasso. The dark blue line depicts the posterior median, the shaded dark blue area depicts the 95% credibility interval. The black dashed lines depict the bootstrapped 95% confidence interval of the classical lasso.

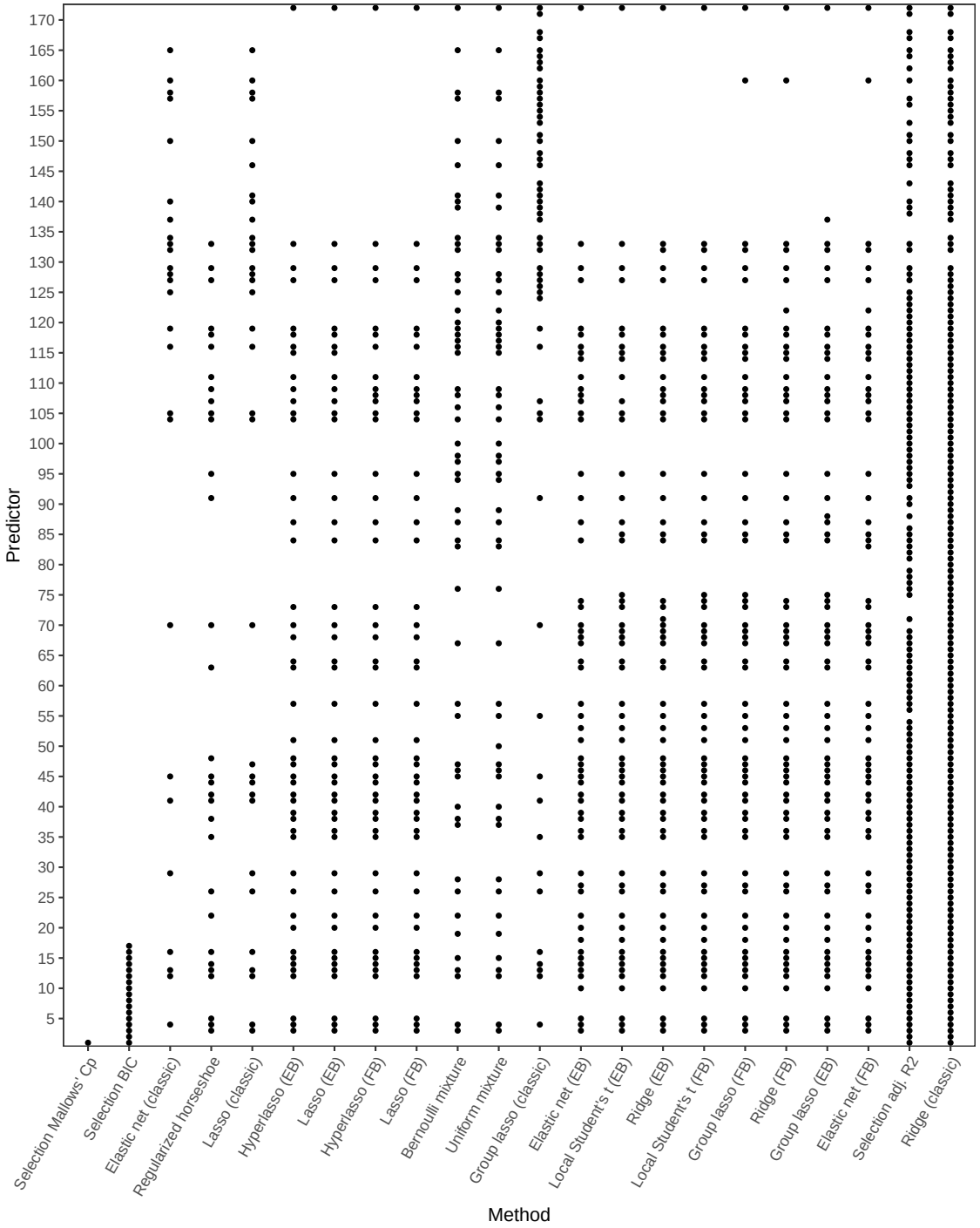

Figure 4.10: Overview of the included predictors for each method in the crime application. Points indicate that a predictor is included based on the optimal credibility interval (CI) from condition 6 in the simulation study. The methods on the x-axis are ordered such that the method that includes the least predictors is on the left and the method that includes the most predictors is on the right.

4.7 Discussion

The aim of this paper was to provide insights about the different shrinkage priors that have been proposed for Bayesian penalization to avoid overfitting of regression models in the case of many predictors. We have reviewed the literature on shrinkage priors and presented them in a general framework of scale mixtures of normal distributions to enable theoretical comparisons between the priors. To model the penalty parameter λ, which is a central part of the penalized regression model, a full Bayes and an empirical Bayes approach were employed.

Although the various prior distributions differ substantially from each other, e.g., regarding their tails or convexity, the priors performed very similarly in the simulation study in those conditions where $p < n$. Overall, the performance was comparable to the classical penalization approaches. The math performance example clearly showed the advantage of using penalization to avoid overfitting when $p < n$. As in the simulation study, the prediction errors in the math example were comparable across penalization methods, although the number of included predictors varied across methods. Finally, although classical penalization is much faster than Bayesian penalization, it does not automatically provide accurate uncertainty estimates and the bootstrapped confidence intervals obtained for the classical methods were generally much smaller compared to the Bayesian credibility intervals.

The differences between the methods became more pronounced when $p > n$. In condition 6 of the simulation study, the (regularized) horseshoe and hyperlasso priors performed substantially better then most of the other shrinkage priors in terms of PMSE. This is most likely due to the fact that the hyperlasso and (regularized) horseshoe are non-convex global-local shrinkage priors and are therefore particularly adept at keeping large coefficients large, while shrinking the small coefficients enough towards zero. Future research should consider various high-dimensional simulation conditions to further explore the performance of the shrinkage priors in such settings, for example by varying the correlations between the predictors. The crime example illustrated the use of the penalization methods further in a $p > n$ situation. In this example, most Bayesian approaches resulted in smaller prediction errors than the classical approaches (except for the mixture priors). Also in terms of the predictors that were included there were considerable differences between the various approaches.

An important goal of the shrinkage methods discussed in this paper is the ultimate selection of relevant variables. Throughout this paper, we have focused on the use of marginal credibility intervals to do so. However, the use of marginal credibility intervals to perform variable selection can be problematic, since the marginal intervals can behave differently compared to joint credibility intervals. This is especially the case for global shrinkage priors, such as the (regularized) horseshoe prior

since these priors induce shrinkage on all variables jointly (Piironen, Betancourt, Simpson, & Vehtari, 2017). Future research should investigate whether the variable selection accuracy can be further improved by using methods that jointly select relevant variables (for example, projection predictive variable selection; Piironen & Vehtari, 2017a, or decoupled shrinkage and selection; Hahn & Carvalho, 2015).

Throughout this paper, we focused on the linear regression model. Hopefully, the results presented in this paper and the corresponding R package `bayesreg` available at `https://github.com/sara-vanerp/bayesreg` will lead to an increased use of penalization methods in psychology, because of the improved performance in terms of prediction error and variable selection accuracy compared to forward subset selection. The shrinkage priors investigated here can be applied in more complex models in a straightforward manner. For example, in generalized linear regression models such as logistic and Poisson regression models, the only necessary adaptation is to incorporate a link function in the model. Although not currently available in the R-package, the available Stan modelfiles can be easily adapted to generalized linear models (GLMs). Additionally, packages such as `brms` (Bürkner, 2017) and `rstanarm` (Stan Development Team, 2016) include several of the shrinkage priors described here, or allow the user to specify them manually. Both packages support (multilevel) GLMs, although `rstanarm` relies on precompiled models and is therefore less flexible than `brms`. Currently, an active area of research employs Bayesian penalization in latent variable models, such as factor models (see e.g., Jacobucci & Grimm, 2018; Lu et al., 2016) and quantile structural equation models (see e.g., Feng, Wang, et al., 2017). The characteristics and behaviors of the shrinkage priors presented in this paper can be a useful first step in solving these more challenging problems.

Chapter 5

A tutorial on Bayesian penalized regression with shrinkage priors for small sample sizes

Based on van Erp, S. (2020). A tutorial on Bayesian penalized regression with shrinkage priors for small sample sizes. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners.*

Abstract

Many of the methods provided in other chapters of this book offer solutions for samples that are small in an absolute sense, for example in single-case designs. In this chapter, the focus is instead on small samples relative to the complexity of the model. I illustrate how Bayesian penalization offers a solution to this problem by applying so-called "shrinkage priors" that shrink small effects towards zero while leaving substantial effects large. A tutorial is provided on applying Bayesian penalization to a linear regression model using the `R` package `bayesreg`, which has various shrinkage priors implemented.

Keywords: Bayesian, Shrinkage Priors, Penalization, Empirical Bayes, Regression.

5.1 Introduction

In the current "Age of Big Data", more and more data is being collected and analyzed. Personal tracking devices allow data to be continuously collected, websites often track online behavior of their users, and large-scale research projects combine data from various sources to obtain a complete picture. These efforts result in large data sets with hundreds or thousands of variables. However, such data sets pose problems in terms of small sample sizes relative to the number of variables. As an example, consider the prediction of the number of murders in a community based on 125 predictors (Redmond & Baveja, 2002). We might use a simple linear regression model to determine the effects of each of the predictors. In order to fit such a model, we would need at least 125 observations, i.e., communities in this case. Now suppose we have collected data on 126 communities. We would be able to fit our linear regression model, but we would be overfitting our model to that specific sample and our results would not generalize well to a different sample from the population (McNeish, 2015). This problem would be exacerbated if we wanted to fit a more complex model including, for example, interactions between the predictors.

Penalization methods offer a solution to this problem. Regular ordinary least squares regression minimizes the sum of squared residuals to find the estimates for regression coefficients. Penalized regression adds a penalty term to this minimization problem. The goal of this penalty term is to shrink small coefficients towards zero, while simultaneously leaving large coefficients large. By doing so, penalization methods aim to avoid overfitting such that the obtained results are generalizable to a different data set from the same population. Popular penalized regression methods include the ridge, lasso, and elastic net penalties. An illustration of the classical lasso penalty is provided in the left column of Figure 5.1. The contours of the sum of squared residuals for two regression coefficients, β_1 and β_2 are shown as black elliptical lines. The classical ordinary least squares solution, $\hat{\beta}_{OLS}$, is the minimum of the sum of squared residuals which lies in the center of the contour lines. The solid black diamond represents the constraint region for the classical lasso penalty function. The lasso solution, $\hat{\beta}_{LASSO}$, is the minimum of the sum of squared residuals plus the lasso penalty term. Graphically, this solution corresponds to the point where the sum of squared residuals contour meets the constraint region of the lasso. It is clear that the lasso solution shrinks both coefficients, with β_1 becoming exactly zero in this example. This illustrates the main advantage of the classical lasso penalty, namely that it can perform automatic variable selection due to its ability to shrink small coefficients to exactly zero. By shrinking the coefficients, penalized regression will lead to an increase in bias but at the same time avoids overfitting (i.e., the bias-variance tradeoff). A comprehensive overview of classical penalized regression can be found in Hastie et al. (2015).

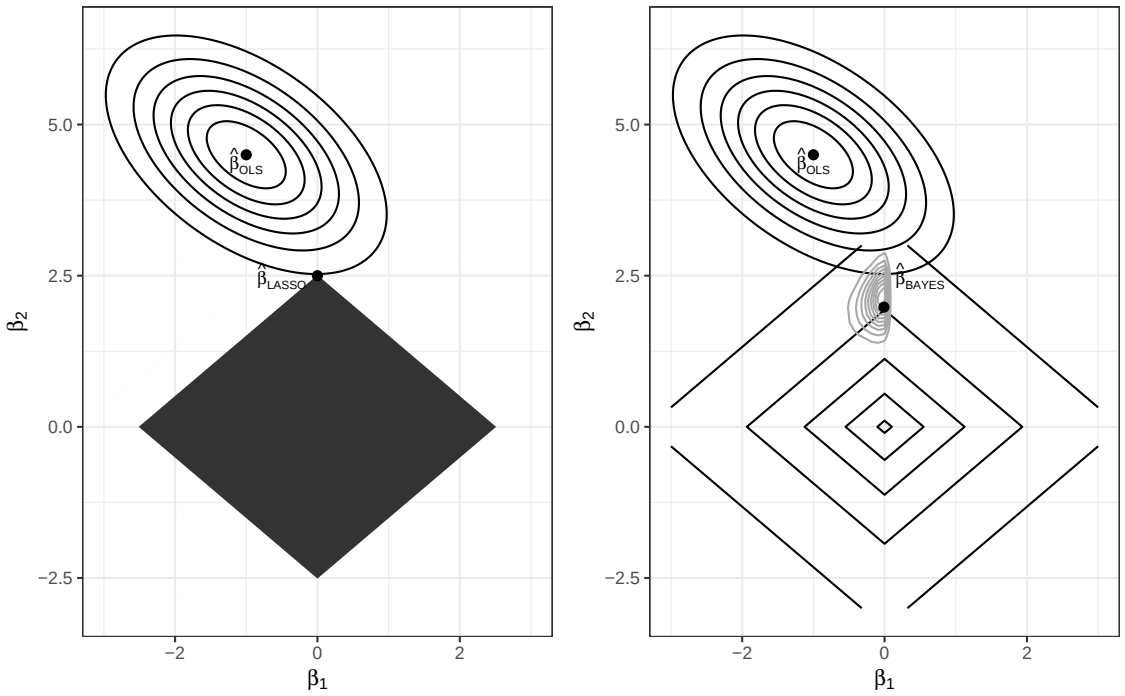

Figure 5.1: Contour plots illustrating classical and Bayesian penalization

The focus of this chapter is on Bayesian penalization, because of several advantages it has over the classical framework. Aside from the usual advantages in terms of automatic uncertainty estimates and intuitive Bayesian interpretations of quantities such as credibility intervals, the Bayesian approach offers three advantages specific to the context of penalization.

5.1.1 Advantage 1. Natural Penalization Through the Prior Distribution

First, penalization can be incorporated naturally in a Bayesian framework through the prior distribution, see also Chapters 1-4 (Miočević, Levy, & Savord, 2020; Miočević, Levy, & van de Schoot, 2020; van de Schoot, Veen, Smeets, Winter, & Depaoli, 2020; Veen & Egberts, 2020). Specifically, we can choose the prior distribution in such a way that it will shrink small effects towards zero, while keeping substantial effects large. By doing so, the prior performs similarly to the penalty term in classical penalized regression. There are prior distributions that, combined with a specific posterior estimate, lead to exactly the same solution as classical penalization methods. For example, specifying double-exponential prior distributions for the regression coefficients will result in posterior modes that are the same as the classical lasso estimates (Park & Casella, 2008). These Bayesian analogues of classical penalization methods have been shown to perform similarly to and in some

cases better than the classical penalization methods (Kyung et al., 2010; Li & Lin, 2010).

5.1.2 Advantage 2. Simultaneous Estimation of the Penalty Parameter

The second advantage of Bayesian penalization lies in the fact that the penalty parameter can be estimated with other model parameters in a single step. The penalty parameter arises in the penalty function of classical penalization methods and determines the amount of shrinkage towards zero. Large values of the penalty parameter lead to more shrinkage towards zero and a penalty parameter equal to 0 will result in no shrinkage at all. Generally, the penalty parameter is determined based on cross-validation, but in Bayesian penalization it is simply a parameter in the prior distribution which can be given its own prior distribution.

5.1.3 Advantage 3. Flexibility in Types of Penalties

The final advantage of Bayesian penalization is that it offers flexibility in terms of the type of penalties that can be considered. Classical penalization methods rely on optimization techniques to find the minimum of the penalized regression function. It is therefore easiest to consider penalty functions that are convex, meaning that they will result in one minimum. Bayesian penalized regression, on the other hand, employs MCMC sampling, which allows a more straightforward implementation of penalties that are not convex.

The right column of Figure 5.1 illustrates Bayesian penalization using the double-exponential prior on the regression coefficients. The elliptical contour lines represent the sum of squared residuals, or the likelihood, centered around the classical ordinary least squares estimate $\hat{\beta}_{OLS}$. The diamond shaped contour lines represent the double-exponential prior, which is similar to the classical lasso constraint region in the left side Figure. The main difference between classical and Bayesian penalization, is the fact that Bayesian penalization results in a full posterior distribution while classical penalization results only in a point estimate. The contour of the posterior distribution is shown in grey and is clearly a compromise between the prior and the likelihood. The posterior mode estimate, $\hat{\beta}_{BAYES}$ is included and corresponds to the classical lasso solution. This double-exponential or lasso prior distribution is just one of many shrinkage priors available. In this chapter, I will summarize the most popular shrinkage priors and illustrate their use in a linear regression model using the flexible software program Stan (Carpenter et al., 2017), see also Chapters 3 (van de Schoot et al., 2020) and 4 (Veen & Egberts, 2020).

5.2 Running example: communities and crime

Throughout this chapter we will use a linear regression model to attempt to predict the number of murders in US communities (Redmond & Baveja, 2002). All code for running this example is available online at the OSF (`osf.io/am7pr/`). The data set is obtained from the UCL machine learning repository (Dua & Graff, 2019) and includes 125 possible predictors (4 are non-predictive and 18 are potential outcomes to predict) of various types of crimes for 2,215 communities. We will focus on the number of murders per 100,000 residents. The predictors include characteristics of the community as well as law enforcement characteristics. Dummy variables are created for the two nominal predictors in the data set, resulting in a total of 172 predictors. All continuous predictors are standardized to have a mean of zero and a variance of one. This is generally recommended in penalized regression to avoid the results depending on the scales of the predictors (Hastie et al., 2015). The implementation of the methods in the `bayesreg` package also requires the predictors to be on the same scale[1].

5.3 Software

There are three different R packages that can be used for Bayesian penalized regression with `Stan`: `rstanarm` (Stan Development Team, 2016), `brms` (Bürkner, 2017), and `bayesreg`. `rstanarm` and `brms` both allow the user to specify multilevel generalized linear models with formula syntax in the same way as classical multilevel generalized linear models are specified in the `(g)lm(er)` functions in `R`. Both packages support various shrinkage priors. The `bayesreg` package is more restricted since it currently only supports linear regression models. Contrary to `rstanarm` and `brms`, the `bayesreg` package is specifically designed to perform Bayesian penalized regression and has all the shrinkage priors implemented that will be discussed in the next section. We will therefore use the `bayesreg` package to illustrate the shrinkage priors in this chapter, although we will note which of the shrinkage priors are available in `rstanarm` and `brms`. All three packages return a `Stan` fit object that can be further processed and several package-specific post-estimation functions.

To fit the Bayesian penalized linear regression model with `bayesreg`, the package needs to be installed first following the instructions available here: `https://github.com/sara-vanerp/bayesreg`. Currently, missing data is not supported in

[1]Throughout this chapter, we use the `bayesreg` package available from `https://github.com/sara-vanerp/bayesreg`. Note that there is also a `bayesreg` package available on CRAN (Makalic & Schmidt, 2016) which has implemented several of the shrinkage priors in linear and logistic regression models. However, contrary to the `bayesreg` package used in this chapter, by Van Erp, the `bayesreg` package on CRAN, by Makalic and Schmidt, only has a subset of the shrinkage priors implemented that are discussed in this chapter.

`bayesreg`. However, it is possible to first impute the missing data using a package such as `mice` (van Buuren & Groothuis-Oudshoorn, 2011) and then fit the model on each of the imputed data sets. The posterior draws for each fitted model can subsequently be combined to obtain the results. For our example, we will simply remove the observations with missingness and focus on the 343 communities with complete data. After installation, the package can be loaded into R and the model can be fit as follows:

```
library(bayesreg)
fit <- stan_reg_lm(X = X, y = y, N_train = 172, prior = "lasso")
```

The required arguments for this function are: a numeric predictor matrix X, a numeric matrix of outcomes Y, the sample size of 172 is used to estimate the model, and the prior choice. The remaining observations in the data are used to estimate the prediction error of the model.

5.4 Shrinkage priors

The goal of a shrinkage prior is to shrink small coefficients towards zero, while keeping large coefficients large. This behavior can be obtained through various types of shrinkage priors, although most shrinkage priors share some general characteristics to ensure this behavior. Specifically, shrinkage priors have a peak at zero to shrink small coefficients. Most shrinkage priors have heavy tails, which allow large coefficients to escape the shrinkage. In this section, we will discuss various shrinkage priors that are popular in the literature. The shrinkage priors are classified into two types: 1) classical counterparts, i.e., shrinkage priors that have been developed as equivalents to classical penalty functions; and 2) Bayesian origin, i.e., shrinkage priors that come from the Bayesian literature and do not have a clear classical counterpart. Table 5.1 provides an overview of the shrinkage priors in each class with references and the `R` packages in which each prior is implemented.

Table 5.1: Overview of the shrinkage priors

Class	Prior	Implemented in	References
Classical counterparts	Ridge	`bayesreg`, `brms`, `rstanarm`	Hsiang (1975)
	Lasso	`bayesreg`, `brms`, `rstanarm`	Park and Casella (2008)
	Elastic net	`bayesreg`	Li and Lin (2010)
Bayesian origin	Student's t	`bayesreg`, `brms`, `rstanarm`	Griffin and Brown (2005); Meuwissen et al. (2001)
	Spike-and-slab	`bayesreg`	George and McCulloch (1993); Mitchell and Beauchamp (1988)
	Hyperlasso	`bayesreg`	Griffin and Brown (2011)
	Horseshoe	`bayesreg`, `brms`, `rstanarm`	Carvalho et al. (2010); Piironen and Vehtari (2017b)
	Regularized horseshoe	`bayesreg`, `brms`, `rstanarm`	Piironen and Vehtari (2017b)

5.4.1 Classical counterparts

Figure 5.2 shows the densities (left) and survival functions (right) for the shrinkage priors corresponding to classical penalty functions. The survival function is equal to 1 minus the cumulative distribution function and is the probability that the parameter has a value greater than the values on the x-axis. For example, at $x = 0$, the survival function equals .5 for all shrinkage priors because each prior is symmetric around zero and thus the probability mass on positive values equals .5. The survival function is insightful to illustrate the tail behavior of the priors: the slower the survival function goes to zero, the heavier the tails. The Bayesian equivalent of the ridge penalty is a normal prior distribution centered around zero. The classical lasso penalty corresponds to a double-exponential prior distribution around zero. It can be seen from Figure 5.2 that the lasso prior is more peaked and has heavier tails compared to the ridge prior. The lasso prior will therefore exert more shrinkage towards zero for small coefficients, but less shrinkage for large coefficients. The classical elastic net penalty is a combination of the ridge and lasso penalties, which becomes apparent from Figure 5.2: its peak and tail lie in between those of the ridge and lasso priors.

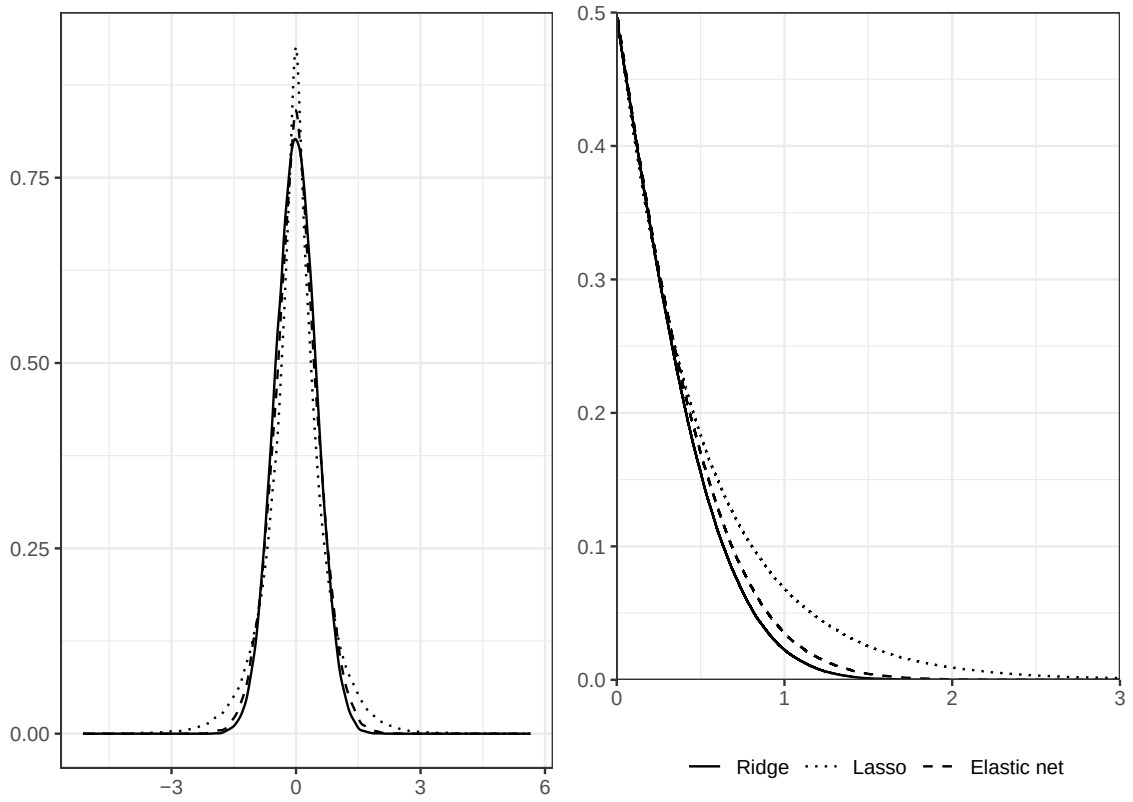

Figure 5.2: Density plot and survival function for the shrinkage priors with a classical counterpart

The exact form of the shrinkage priors depends on the values of the hyperpa-

rameters in the priors. For the ridge and lasso priors, the only hyperparameter is the scale which influences how spread out the prior will be. In `bayesreg`, these scales are equal to $\frac{\sigma_\epsilon}{\lambda}$, where σ_ϵ is the standard deviation of the errors. Especially for the lasso prior, including the error standard deviation in the prior is important to avoid multimodal posteriors (Park & Casella, 2008). The λ parameter has a similar role to the penalty parameter in classical penalized regression. Larger values for λ result in a smaller prior variance and thus more shrinkage towards zero. The elastic net prior requires specification of two penalty parameters: λ_1 which determines the influence of the lasso, and λ_2 which determines the influence of the ridge. Thus, setting λ_1 to 0 results in the ridge prior and setting λ_2 to zero results in the lasso prior. In `bayesreg`, the λ parameter is given a standard half-Cauchy prior distribution, so that its value is automatically determined by the data. However, other options to determine λ are possible, such as empirical Bayes methods or cross-validation.

5.4.2 Bayesian origin

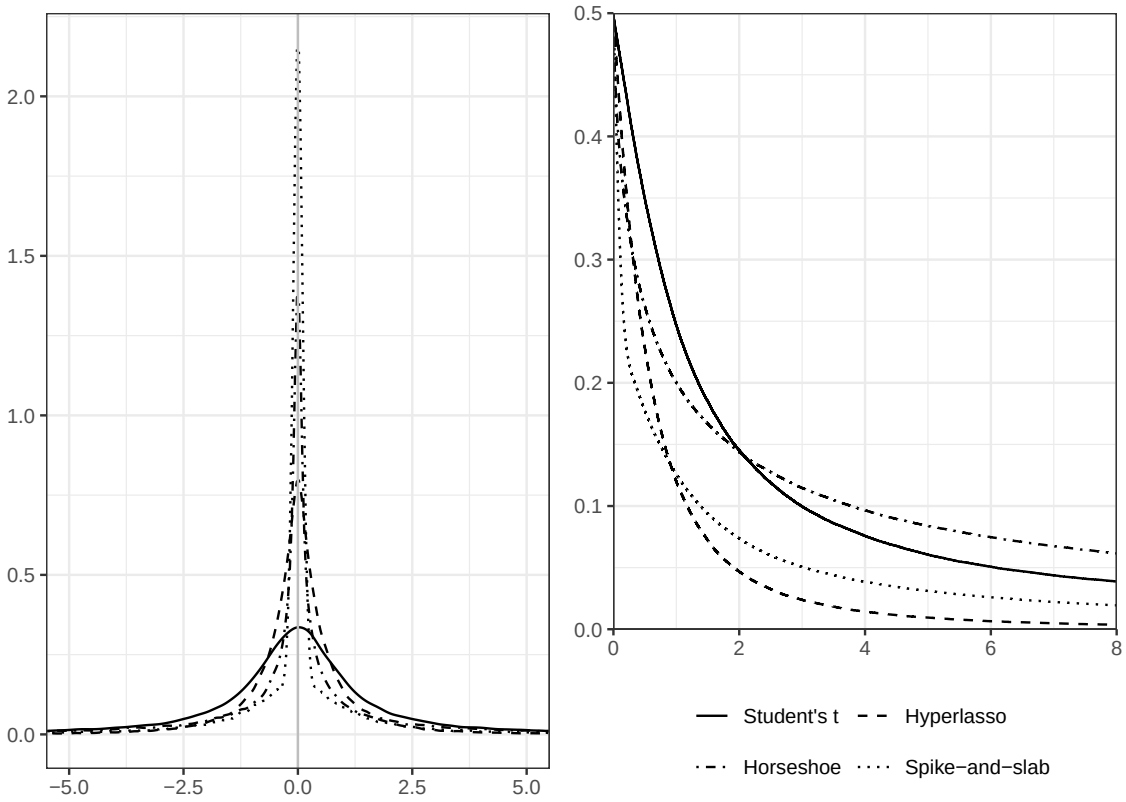

Figure 5.3: Density plot and survival function for the shrinkage priors with a Bayesian origin

Figure 5.3 presents the densities (left) and survival functions (right) for the shrinkage priors with a Bayesian origin. Student's t-distribution is similar to a normal distribution but has heavier tails. As a result, Student's t prior is more adept

at leaving substantial coefficients large compared to the ridge prior. However, Student's t prior is not as peaked around zero compared to the other shrinkage priors with a Bayesian origin. The more peaked the distribution, the more shrinkage towards zero for small coefficients. The hyperlasso prior is more peaked around zero. The hyperlasso can be seen as an extension of the lasso, but with heavier tails to avoid too much shrinkage of large coefficients. The hyperlasso is a so-called global-local shrinkage prior, with a global shrinkage parameter that simultaneously shrinks all coefficients towards zero and local shrinkage parameters for each regression coefficient that allow large coefficients to escape the global shrinkage. The horseshoe prior is another global-local shrinkage prior that is very popular in the Bayesian literature. It has an asymptote at zero and heavy tails, which make the horseshoe very adept at shrinking small coefficients heavily towards zero but leaving the large coefficients large. In practice, however, it might be necessary to have some shrinkage of large coefficients. For example, some parameters might be weakly identified meaning that there is not enough information in the data to estimate them. In this case, the heavy tails of the horseshoe prior can lead to an unstable MCMC sampler. The regularized horseshoe prior has been proposed to solve this issue. Its density is not included in Figure 5.3 since it is very similar to that of the horseshoe. Specifically, small coefficients will be shrunken in the same way as with the horseshoe prior. The main difference is that the regularized horseshoe induces some slight shrinkage on large coefficients as well. Finally, we have the spike-and-slab prior which is a mixture of two distributions: a peaked distribution around zero for the small coefficients (the spike), and a vague distribution for the large coefficients (the slab). The `bayesreg` implementation of the spike-and-slab prior has a normal spike with a very small variance of .001, which is very peaked around zero and a Cauchy slab, which has heavy tails. This can also be seen from Figure 5.3.

The hyperparameters that need to be specified for each of the shrinkage priors with a Bayesian origin in `bayesreg` vary. For the Student's t and horseshoe priors, no hyperparameters need to be specified since all parameters are given a prior distribution in the program. For the hyperlasso, the degrees of freedom need to be specified with smaller degrees of freedom resulting in a heavier-tailed prior. The default value in bayesreg is .5, but similar to the horseshoe prior, this might not shrink weakly identified parameters enough so it might be necessary to specify a higher value for the degrees of freedom. The regularized horseshoe prior has the most flexibility in terms of tuning. First, for the global shrinkage parameter (which determines the general shrinkage for all coefficients simultaneously) a scale (`scale_global`) and degrees of freedom (`global_df`) parameter need to be specified. The scale influences how wide the peak is and defaults to 1. A smaller scale leads to more overall shrinkage of all coefficients. If prior information regarding the

number of relevant predictors is available, it is better to determine the global scale based on this information. This can be done by setting the `p0` argument equal to the a priori assumed number of relevant predictors. The global degrees of freedom parameter determine the tail behavior and defaults to 1, with larger values leading to lighter tails. For the local shrinkage parameters (which allow truly large coefficients to escape the global shrinkage), only the degrees of freedom (`local_df`) need to be specified, with 1 as default and larger values resulting in lighter tails. Finally, the regularized horseshoe differs from the horseshoe prior by asserting some shrinkage on large coefficients. This shrinkage is determined by a t-distribution with some scale (`slab_scale`) and degrees of freedom (`slab_df`). Both default to 1. Finally, for the spike-and-slab prior, a decision needs to be made on the prior for the mixing probabilities. The mixing probabilities influence whether a coefficient falls in the spike or the slab of the prior, and thus whether the coefficient will be shrunken heavily towards zero (in case of the spike) or not (in case of the slab). The first option in `bayesreg` is a Bernoulli prior on the mixing probabilities, in which each coefficient will be assigned to either the spike or the slab, with probability .5. The second option is a uniform prior, which is more flexible since the prior on each coefficient will be a mixture of the spike and the slab, where the influence of the spike and the slab is weighted by the mixing probabilities.

5.5 Practical considerations

So far, we have discussed various shrinkage priors. However, in order to apply these shrinkage priors, there are some practical issues to consider. These issues include: 1) how to choose a shrinkage prior; and 2) how to select variables based on the results.

5.5.1 Choice of the shrinkage prior

The type of prior information encoded in shrinkage priors is the same: some of the values for the coefficients are so small, they should be shrunken towards zero, and only substantial coefficients should remain large. However, the priors vary in the way this information is translated in practice. First, depending on the prior used and the hyperparameters chosen, the amount of shrinkage towards zero for small coefficients varies. In general, the more peaked the prior is around zero, the heavier the shrinkage for small coefficients. Second, the amount of shrinkage for large coefficients varies across priors and hyperparameters. This is mainly influenced by the heaviness of the tails. For example, compared to the lasso prior, the ridge prior has lighter tails and will therefore shrink large coefficients more towards zero than the lasso prior (given that the scale is the same in both priors). The first

step in choosing a specific shrinkage prior and its hyperparameters is therefore to understand its behavior. This can be easily done by sampling draws from various priors and hyperparameter settings and comparing the density plots. To this end, the code for creating Figures 5.2 and 5.3 are made available online at `osf.io/am7pr/` and can be adapted to compare various hyperparameter settings.

In general, the goal of Bayesian penalization is to avoid overfitting. To evaluate this property, we can split the data in a training and test set. We estimate the model on the training set and then use the resulting estimates for the regression coefficients to compute the responses in the test set. The prediction mean squared error (PMSE) summarizes the prediction error by taking the mean of the squared differences between computed and true responses in the test set. In `bayesreg`, the function `pmse_lm` computes the PMSE. In general, when the number of predictors is smaller than the sample size, most shrinkage priors discussed in this chapter will lead to similar prediction errors. The shrinkage priors vary more in terms of prediction errors when the number of predictors exceeds the sample size. There is some evidence that global-local shrinkage priors such as the (regularized) horseshoe and hyperlasso perform best in this situation (van Erp, Oberski, & Mulder, 2019), but more research in this area is required. One option to choose a shrinkage prior for the application at hand is to fit the model using various shrinkage priors and then use the PMSE to guide the choice for the prior. When reporting the results, it is important to describe that this strategy was used and which other shrinkage priors (including their hyperparameters) were considered.

There are two other important criteria to consider when choosing a shrinkage prior: 1) computation time, and 2) desired complexity of the resulting model. First, the computation time can vary greatly between the shrinkage priors. In general, if a shrinkage prior becomes more complex, the computation time increases, especially when adaptation of the HMC sampler settings is needed. Second, since the shrinkage priors vary in the amount of shrinkage they perform, the eventual number of excluded predictors can vary across shrinkage priors. Thus, if a very sparse solution is desired, a very peaked shrinkage prior should be chosen. Note that the number of excluded predictors depends heavily on the criterium that is used to select predictors, which will be discussed in the next subsection.

To continue with the communities and crime example, let us compare several shrinkage priors according to the criteria mentioned above. Recall that we have a total of 172 predictors (including recoded dummy variables) and observations from 343 communities. Half of the observations (172) are used as training set and the remaining 171 observations are used to test the model. Three different shrinkage priors are compared: the lasso, the hyperlasso, and the spike-and-slab prior with Bernoulli mixing probabilities. We can fit for example the spike-and-slab prior as

follows:

```
fit.ssp <- stan_reg_lm(X = X, y = y, N_train =172, prior = "mixture",
hyperprior_mix = "Bernoulli", iter = 2000, chains = 4, seed = 27022019)
```

The spike-and-slab prior takes longest with 367 seconds and results in 3851 transitions after warmup that exceeded the maximum treedepth. Therefore, we need to increase the `max_treedepth` setting of the sampler above 10, which will lead to an increased computation time. In general, the spike-and-slab prior has a large computation time and we might decide to not choose this prior based on time considerations. The computation time is lowest for the hyperlasso (20 seconds), followed by the lasso (29 seconds). The PMSEs for the lasso and hyperlasso do not differ much (55.4 and 55.2, respectively). Note there are no clear cutoffs out there when a difference in PMSE is substantial or not, so it comes down on personal interpretation.

5.5.2 Variable selection

One of the main goals of penalized regression is to automatically select relevant predictors. Classical penalization methods such as the lasso are able to shrink small coefficients exactly to zero, thereby performing automatic variable selection. Bayesian penalization methods, on the other hand, do not perform automatic variable selection and thus a criterion is needed to select relevant predictors. Different criteria exist. One option is to simply include those predictors for which the posterior estimate exceeds a certain cut-off value, such as .1. However, it has been shown that this arbitrary choice of cut-off value leads to high false inclusion rates (van Erp et al., 2019). A second option is to include a predictor when the credibility interval for that coefficient does not cover zero. In this approach, a choice needs to be made regarding the posterior probability to include in the credibility interval. The optimal credibility interval in terms of correct and false inclusion rates varies across priors and types of data sets. An overview of optimal credibility intervals for various simulated data sets can be found in van Erp et al. (2019) and I use this overview to determine the credibility interval to use for the communities and crime example. Since we have 172 predictors and 172 observations in the training set, we select the optimal credibility intervals corresponding to condition 6, in which the ratio of predictors to observations is most equal to our example, leading to 30% intervals for both priors. We can then use the following function in bayesreg to select the variables:

```
select_lm(fit, X = X, prob = 0.3)
```

In this case, the priors select almost the same number of variables, 50 for the lasso and 47 for the hyperlasso. It appears that the shrinkage priors perform very similarly in this application, both in terms of prediction error and in terms of variable selection. To check this graphically, we can plot the posterior estimates and credibility intervals for the priors. Here, we will use the 30% credibility intervals, to immediately see which predictors are included in the model:

```
fitlist <- list(fit.lasso, fit.hyperlasso)
names(fitlist) <- c("lasso", "hyperlasso")
plots <- plot_est(fitlist, est = "mean", CI = 0.30, npar = 50, pred.nms
= colnames(X))
```

The function returns a list of plots such as the one presented in Figure 5.4 Indeed, we see no substantial differences between the results of the lasso and hyperlasso. Of the predictors shown in Figure 5.4, both shrinkage priors select the racial match between community and police force, the number of police officers, and the per capita income as predictors for the number of murders in the community.

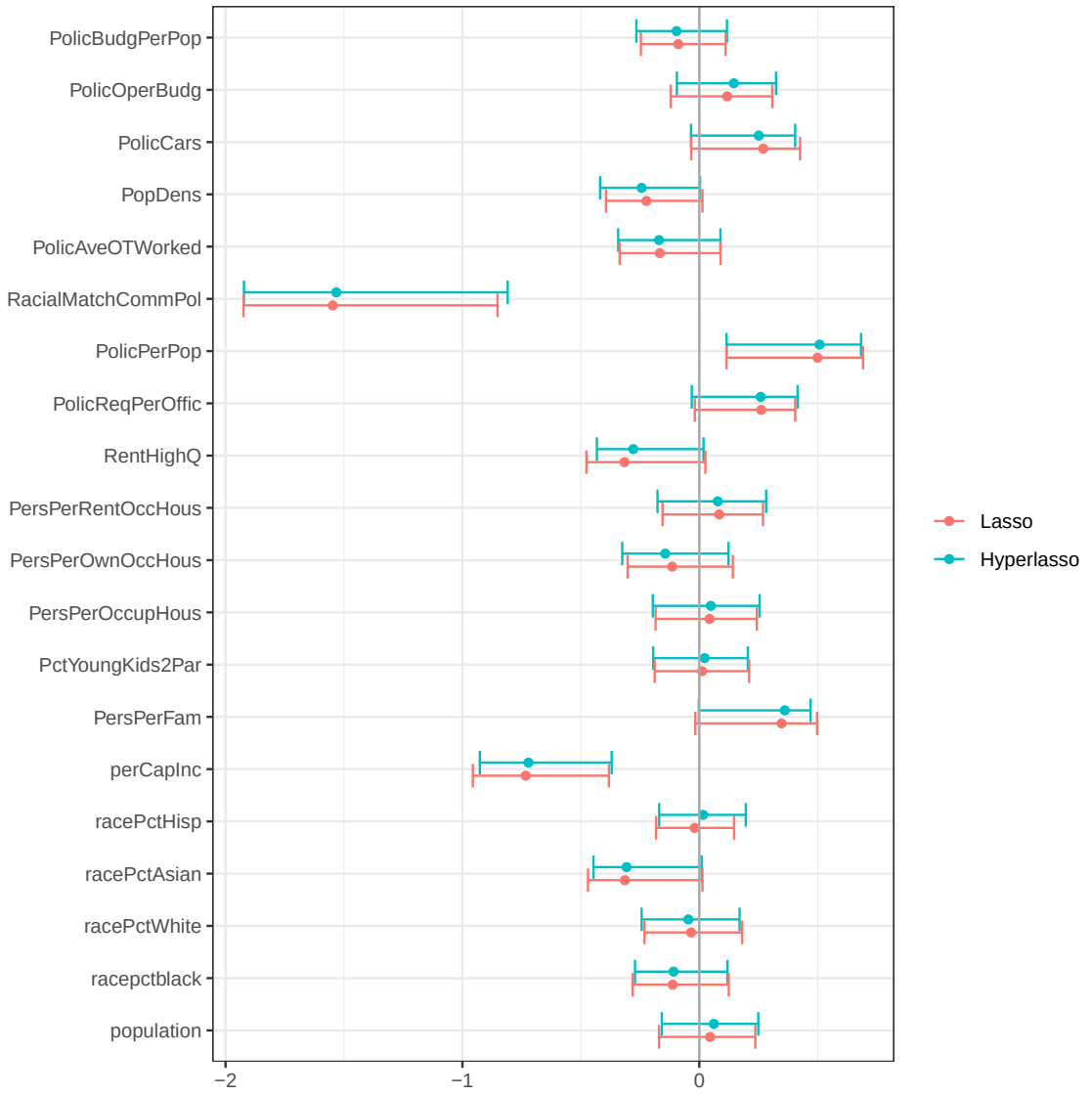

Figure 5.4: Comparison of posterior mean estimates and 30% credibility intervals obtained with the lasso and hyperlasso priors for a selection of predictors

One issue with the credibility interval criterion for variable selection is its dependency on the posterior probability included in the interval, which differs across shrinkage priors and data characteristics. Moreover, credibility intervals only consider the marginal posteriors per regression coefficient separately. This might not be optimal for shrinkage priors that shrink parameters jointly (e.g., the global-local shrinkage parameters), in which case the joint credibility interval might perform differently than the marginal intervals (Piironen et al., 2017; van der Pas, Szabó, & van der Vaart, 2017). An alternative that does take into account the joint posterior distribution is projection predictive variable selection which is implemented in the `projpred` package (Piironen & Vehtari, 2015).

Chapter 6

Shrinkage priors for Bayesian measurement invariance: A robust approach for modeling and detecting non-invariance

Based on van Erp, S., Mulder, J., and Oberski, D.L. (In preparation). Shrinkage priors for Bayesian measurement invariance: A robust approach for modeling and detecting non-invariance.

Abstract

Measurement invariance (MI) is of central importance whenever latent constructs are being compared across groups. Traditional approaches to model measurement invariance often rely on unrealistic model restrictions followed by post-hoc modifications. More recently, Bayesian approximate MI and the alignment method have been proposed to avoid the use of such unrealistic restrictions. In this paper, we propose a novel method to further improve on these state-of-the-art approaches. We rely on the use of robust shrinkage priors to automatically model measurement invariance. Robust shrinkage priors are heavily peaked around zero with thick tails, allowing them to identify the model while simultaneously enabling noninvariant measurement parameters to escape the shrinkage. Specifically, we focus on the spike-and-slab prior and the regularized horseshoe prior. We discuss how the exact and approximate invariance approaches can be viewed as specific types of shrinkage priors and we compare them to the more robust spike-and-slab and regularized horseshoe prior. We show how these more robust shrinkage priors outperform the alignment method and approximate MI in terms of factor mean estimation when large amounts of noninvariance are present. Finally, we apply the shrinkage priors to data from the European Social Survey and we illustrate how the results might be used to assess which measurement parameters show violations of invariance.

Keywords: Measurement Invariance, Bayesian, Shrinkage Priors.

6.1 Introduction

Measurement invariance (MI) is an important concept in any study where latent constructs are being compared across groups, countries, or time-points. From a conceptual point of view, MI implies that the latent construct under investigation is measured in the same way across groups or time-points, while measurement non-invariance (MNI) indicates that the measurement instrument behaves differently in certain groups or on certain time-points. In the case of MNI, comparisons across groups or time points become more complicated because differences in item scores between individuals of different populations can either be caused by true differences between the populations or by varying interpretations of the items.

Generally, multiple group confirmatory factor analysis (MGCFA) is used to compare latent constructs across groups or over time. The focus of such analyses lies mainly in comparing the factor means and variances. For example, cross-country comparative surveys such as the European Social Survey (ESS) or the World Values Survey (WVS) measure and compare different latent constructs such as values or beliefs across countries. Surveys such as the World Health Survey (WHO) focus on measuring health and health-related topics across countries. Achievement surveys, such as the Programme for International Student Assessment (PISA), test students' latent abilities in order to evaluate and compare educational systems across countries. In addition to these cross-sectional surveys which are conducted repeatedly, panel surveys measure the same people multiple times. For example, the European Community Household Panel (ECHP) repeatedly interviews a panel on topics related to living conditions. Instead of comparing latent constructs across groups, such longitudinal studies aim to compare latent constructs over time.

In order to make any valid comparisons, restrictions need to be imposed on the model to ensure that the latent construct under investigation has a similar meaning across the groups. We can distinguish two types of restrictions in the multiple group factor model: 1) restrictions to identify the scale of the latent variable in one group; and 2) restrictions to link the scale of the latent variable across groups. The first type of restrictions is needed in any factor model and can be achieved by fixing the mean of the factor or fixing one intercept to zero in combination with fixing the variance of the factor or fixing one loading to one. The second type of restrictions ensures that the measurement model is invariant across the groups. Traditionally, equivalency restrictions are imposed across the groups, starting with the loadings (metric invariance) and then adding the restriction that the intercepts should be equal across groups (scalar invariance), and possibly continuing with the factor covariances and error variances (Steenkamp & Baumgartner, 1998). At each level of invariance, the fit of the model is assessed. In the case of many groups, full metric invariance can already be hard to obtain, while full scalar invariance is practically impossible.

Modification indices are often used to determine which parameter restrictions might be freed to instead attain partial metric and/or scalar invariance (Byrne, Shavelson, & Muthén, 1989). In the case of many groups, the two main problems with this approach, however, are: 1) due to the many post-hoc adaptations to the model, the final model might be the results of chance capitalization (MacCallum, Roznowski, & Necowitz, 1992); and 2) it is inefficient and time-consuming since the model needs to be re-estimated after each restriction has been freed.

To solve these problems, two approaches have been proposed that do not rely on modification indices. The first approach is Bayesian approximate MI (B. O. Muthén & Asparouhov, 2013). In this approach, a normal prior distribution is specified for each pairwise difference between measurement parameters. By specifying a small variance for the normal prior, the measurement parameters are restricted to be approximately equal across groups, instead of exactly equal as in the traditional approach. Inferences are then made based on the posterior distribution, which combines the information in the prior distribution with the information in the data. There are multiple advantages to using a Bayesian approach. First, it is possible to incorporate prior knowledge in the analysis. In the case of approximate MI, the prior knowledge that is included states that the measurement parameters are approximately equal across groups, instead of exactly equal, which seems more plausible in practice. Second, Bayesian methods provide a natural way to impute missing data (see e.g., Gelman et al., 2013, Chapter 18). Third, the results are more straightforward to interpret. For example, 95% credibility intervals can be computed and interpreted as the interval in which the true value lies with 95% probability (see e.g., Berger, 2006). Disadvantages of Bayesian approximate MI, however, are: 1) the results can be greatly influenced by the arbitrarily chosen prior variance (which quantifies the degree of noninvariance; B. O. Muthén and Asparouhov (2013); van de Schoot et al. (2013)); and 2) the approach is not robust in the sense that only small deviations from MI can be modeled. If there are a few highly non-invariant parameters, the small variance priors will lead to bias in the estimates (van de Schoot et al., 2013). Therefore, a two-step approach is needed when applying Bayesian approximate MI in which non-invariant parameters are first identified using the small variance priors and then freed in a second analysis (B. O. Muthén & Asparouhov, 2013). Thus, Bayesian approximate MI does lessen the problems associated with modification indices, but it does not fully solve them.

The second approach to modeling noninvariance without needing to rely on modification indices is the alignment method (Asparouhov & Muthén, 2014). This is a two-step procedure that first fits a base model, such as the configural model with the factor means and variances fixed to 0 and 1 in each group. Next, a simplicity function is minimized with respect to the factor means and variances to obtain

a solution in which the number of non-invariant items is minimized. Asparouhov and Muthén (2014) present both a maximum likelihood and a Bayesian approach for the alignment method. In the Bayesian approach, the base model can be either the configural model or the approximate MI model with all measurement parameters approximately equal and only the factor means and variances in the first group fixed. The simplicity function is then minimized in each Markov Chain Monte Carlo (MCMC) iteration. Compared to Bayesian approximate MI, the Bayesian alignment method is more robust in the sense that some items are allowed to have large violations from MI. Nevertheless, the alignment method still assumes only a small to moderate amount of non-invariance (i.e., up to 25% of the items) and has been shown to perform badly when there are many large violations of MI (Asparouhov & Muthén, 2014; Flake & McCoach, 2018). Moreover, the alignment method does not allow the user to incorporate prior knowledge, apart from the choice of the base model (i.e., configural or approximate MI). Especially in large scale surveys which are performed every few years, knowledge regarding the amount of non-invariance of certain scales might be available from previous rounds. Asparouhov and Muthén (2014) proposed an ad-hoc procedure to determine which measurement parameters are (approximately) invariant. In a simulation study, Flake and McCoach (2018) showed that this approach has low power to flag non-invariant items, especially when itemscores are skewed. Finally, the alignment method can currently only be applied in models without cross-loadings or covariates.

In this paper, we propose to extend the current toolbox for modeling and assessing MI using so-called robust shrinkage priors. We refer to shrinkage priors as any Bayesian prior distribution that is not flat, since it will exert some shrinkage on the estimates. However, in this paper we focus specifically on what we refer to as robust shrinkage priors. Robust shrinkage priors are heavily peaked around zero with thick tails. As a result, these priors heavily shrink small deviations from MI towards zero, while exerting no or almost no influence on large deviations from MI. The methods are Bayesian, thus offering the advantages of the Bayesian framework in terms of missing data handling and interpretability. Shrinkage priors are popular in linear regression problems with many predictors since they aim to shrink small coefficients towards zero, thereby preventing overfitting (van Erp et al., 2019). More recently, shrinkage priors have been applied in latent variable models (see e.g., Feng, Wang, et al., 2017; Jacobucci & Grimm, 2018; Lu et al., 2016), although not yet in the multiple group confirmatory factor model.

We will focus on two state-of-the-art shrinkage priors for modelling MI: the spike-and-slab prior and the regularized horseshoe prior. The spike-and-slab prior has recently been applied in confirmatory factor models (Lu et al., 2016) as well as in the three parameter logistic (3PL) model to model differential item functioning

(DIF; Soares, Gonçalves, & Gamerman, 2009). The regularized horseshoe prior has only been applied in regression models (Piironen & Vehtari, 2017b). Both priors provide a more robust alternative to approximate MI and the alignment method since both small and large deviations from MI are automatically modelled. Additionally, the proposed methods are expected to improve upon the (Bayesian) alignment method because they allow the user to easily incorporate prior knowledge regarding the number of non-invariant measurement parameters and they can be applied in situations with more than 25% MNI as well as in more complex models with cross-loadings and/or covariates.

The remainder of this paper is organized as follows: Section 6.2 presents the Bayesian multiple group confirmatory factor model. Section 6.3 describes and compares possible prior distributions to model measurement invariance, based on commonly used methods as well as the newly proposed shrinkage priors. Section 6.4 illustrates the behavior of the shrinkage priors in a small simulation study and the priors are applied to empirical data in Section 6.5. Finally, the results are discussed in Section 6.6.

6.2 The Bayesian multiple group confirmatory factor model

6.2.1 Multiple group confirmatory factor model

The multiple group factor model for continuous items y_{ijg} loading on a single factor η_{ig} is given by:

$$\begin{aligned} y_{ijg} = \nu_{jg} + \lambda_{jg}\eta_{ig} + \epsilon_{ijg}, \quad & \text{with } \eta_{ig} \sim N(\alpha_g, \omega_g^2), \\ & \text{and } \epsilon_{ijg} \sim N(0, \sigma_{jg}^2), \end{aligned} \tag{6.1}$$

for $i = 1, \ldots, n_g$ individuals, $j = 1, \ldots, J$ items, and $g = 1, \ldots, G$ groups.

The measurement intercepts and factor loadings in Model 6.1 can be reparametrized in terms of deviances from the average value over the groups for each measurement parameter, i.e. $\delta_{jg}^{\nu} = \nu_{jg} - \mu_j^{\nu}$ and $\delta_{jg}^{\lambda} = \lambda_{jg} - \mu_j^{\lambda}$. The multiple group factor model then becomes:

$$\begin{aligned} y_{ijg} = (\delta_{jg}^{\nu} + \mu_j^{\nu}) + (\delta_{jg}^{\lambda} + \mu_j^{\lambda})\eta_{ig} + \epsilon_{ijg}, \quad & \text{with } \eta_{ig} \sim N(\alpha_g, \omega_g^2), \\ & \text{and } \epsilon_{ijg} \sim N(0, \sigma_{jg}^2). \end{aligned} \tag{6.2}$$

Throughout this paper, we will work with the second parametrization.

In a Bayesian analysis, a prior distribution is specified for each parameter in the model. Thus, for the multiple group factor model we have the joint prior distribution $p(\delta_{jg}^{\nu}, \mu_j^{\nu}, \delta_{jg}^{\lambda}, \mu_j^{\lambda}, \alpha_g, \omega_g^2, \sigma_{jg}^2) = p(\delta_{jg}^{\nu})p(\delta_{jg}^{\lambda})p(\mu_j^{\nu})p(\mu_j^{\lambda})p(\alpha_g)p(\omega_g^2)p(\sigma_{jg}^2)$. We specify independent priors for the model parameters, which is a standard default choice. Only for the regularized horseshoe, which is discussed in Section 6.3.5, the priors for the deviance parameters are conditional on the error variances σ_{jg}, i.e., $p(\delta_{jg}^{\nu}|\sigma_{jg})$ and $p(\delta_{jg}^{\lambda}|\sigma_{jg})$.

The focus of this paper lies on the priors for the deviance parameters $p(\delta_{jg}^{\nu})$ and $p(\delta_{jg}^{\lambda})$, because these parameters represent the violations from MI. Specifically, larger values for δ_{jg}^{ν} and δ_{jg}^{λ} imply greater violations of MI. By specifying heavy-tailed shrinkage priors for these parameters, we aim to shrink the small, negligible violations from MI towards zero while allowing the large, substantially important violations to remain large.

For the other parameters in the model, we specify the following weakly informative priors:

$$\begin{aligned} p(\mu_j^{\nu}) &\sim \text{Normal}(0, 100) \\ p(\mu_j^{\lambda}) &\sim \text{Normal}(0, 100) \\ p(\alpha_g) &\sim \text{Normal}(0, 100) \\ p(\omega_g) &\sim \text{half-Cauchy}(0, 5) \\ p(\sigma_{jg}) &\sim \text{half-Cauchy}(0, 5) \end{aligned} \tag{6.3}$$

Note that many other weakly informative prior options are possible. Moreover, if prior information is available, informative priors can be specified instead. It is always recommended to perform a sensitivity analysis to assess the sensitivity of the results to the specific choice of the prior (van Erp et al., 2018).

After the prior distributions have been specified and the data has been collected, the likelihood of the data and the joint prior distribution are multiplied to obtain the posterior distribution, i.e.,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{6.4}$$

with $p(\boldsymbol{\theta})$ containing all parameters in the model. Draws from the posterior distribution are generally obtained through Markov Chain Monte Carlo (MCMC) sampling. We will now discuss the necessary identification restrictions for the multiple group factor model.

6.2.2 Identification restrictions

As mentioned in the Introduction, we can distinguish between two types of restrictions to identify the model: 1) restrictions to identify the factor in one group; and 2) restrictions to link the latent variable scale across groups. Throughout this paper, we will identify the factor in one group by fixing the factor mean in the first group α_1 to 0 and by fixing the average loading for item 1 across groups μ_1^λ to 1. Alternatively, we can fix the factor variance in one group to 1. However, this approach results in bimodality due to sign changes in the loadings between the chains (Erosheva & Curtis, 2017).

If we were to impose a vague normal prior on the deviance parameters δ_{jg}^ν and δ_{jg}^λ, the model would not be identified because such a vague prior allows too many of the measurement parameters to differ between the groups. Thus, we need to impose a prior on the deviance parameters that results in heavy shrinkage near zero so that the latent variable scale is linked across the groups and the latent variable means and variances are identified. Ideally, we also need the prior to not, or hardly, shrink truly large deviances so that violations of MI are automatically accounted for within the model. We will use shrinkage priors for this purpose. Note that we will specify the shrinkage priors for the deviance parameters in all but the first group and we impose a sum-to-zero constraint on the deviances for each item in the first group, i.e., $\delta_{j1}^\nu = -\Sigma_{g=2}^G \delta_{jg}^\nu$ and $\delta_{j1}^\lambda = -\Sigma_{g=2}^G \delta_{jg}^\lambda$. Next, we discuss various prior distributions that can be specified in the context of measurement invariance.

6.3 Prior distributions to model measurement invariance

We start by placing three common methods to model measurement invariance - exact invariance, approximate invariance, and alignment - within the Bayesian framework before turning to the proposed robust shrinkage priors. By viewing all approaches in light of the prior distributions they imply, we facilitate comparison of the various methods. The methods can be seen as moving from most restrictive and least robust to least restrictive and most robust.

Currently, the alignment method is the most robust method available to model MI. However, because of the small number of violations of MI that can escape the shrinkage in the alignment method (i.e., up to 25% of the items), it is of interest to consider a fully Bayesian method relying on robust, heavy-tailed priors. This enables us to investigate whether we can allow a larger number of substantial deviations from MI while still obtaining an identified model. Specifically, we consider two shrinkage priors that are popular in the Bayesian literature: the spike-and-slab and

the regularized horseshoe prior. They have in common a sharp peak around zero and heavy tails. The sharp peak pulls small deviations towards zero, thereby identifying the model, while the heavy tails allow large deviations to escape the shrinkage, thereby avoiding biased estimates.

6.3.1 Exact invariance

Traditionally, the most popular method to establish MI imposes exact equality restrictions on the measurement parameters and relies on fit measures to assess whether these restrictions hold. From a Bayesian viewpoint, this approach is equivalent to imposing a prior consisting of a point mass at zero for the pairwise deviances between measurement parameters or the deviances δ_{jg}^{ν} and δ_{jg}^{λ}.

6.3.2 Approximate invariance

Instead of the exact equality constraints, approximate MI allows some "wiggle room" in the measurement parameters. It does so by setting normal priors on the pairwise differences between measurement parameters with a mean of zero and a small variance, such as 0.001[1]. As a result, pairwise differences between the measurement parameters are shrunken towards zero. Note that this is different from specifying a small-variance normal prior on the deviances δ_{jg}^{ν} and δ_{jg}^{λ}, which would shrink the measurement parameters towards the average across groups. Regardless of the parametrization, approximate MI is not a robust approach. The results depend greatly on the arbitrarily chosen prior variance of the pairwise differences and it is assumed that there are many small deviations from MI. Large deviations from MI will be shrunken heavily towards zero, thereby biasing estimates of interest such as the factor means.

6.3.3 Alignment

Alignment relies on the minimization of a simplicity function. As noted by Asparouhov and Muthén (2014) this use of a simplicity function is similar to rotation in exploratory factor analysis (EFA), which in turn has been compared to the use of small-variance priors to identify structural equation models (Guo et al., 2019). The alignment method aims to find a solution in which most measurement parameters are approximately equal while only a few measurement parameters show large violations from MI. It is currently unknown which prior distribution exactly corresponds to

[1]Note that the implementation in Mplus uses multivariate normal prior distributions on the measurement parameters with the hyperparameters chosen such that the implied prior for the pairwise differences is centered around zero with a small variance (see Appendix 9.1 in B. O. Muthén & Asparouhov, 2013).

the alignment method. However, it can be expected that this prior will have heavier tails than the normal distribution to allow a small number of large violations to escape the shrinkage towards zero.

6.3.4 Robust invariance: the spike-and-slab prior

The spike-and-slab prior is a discrete mixture of a peaked distribution or point mass around zero (the spike) and a vague proper prior (the slab). First proposed, among others, by Mitchell and Beauchamp (1988), the spike-and-slab prior has become a popular prior in the Bayesian literature and various formulations exist (see e.g., George & McCulloch, 1993; Ishwaran & Rao, 2005). More recently, the spike-and-slab prior has been applied in the context of confirmatory factor analysis to determine which cross-loadings should be included in the model (Lu et al., 2016). However, the spike-and-slab prior proposed by Lu et al. (2016) was not used to identify the factor model, only to detect the relevant cross loadings after minimal identification restrictions were included in the model. In the context of MI, the spike-and-slab prior has been used by Soares et al. (2009) in the three parameter logistic model. In this case, the spike-and-slab prior was used to identify the model. We adapt the spike-and-slab prior for the multiple group confirmatory factor model.

We use the following formulation of the spike-and-slab prior based on George and McCulloch (1993):

$$\begin{aligned} \delta_{jg}|\pi_{jg}, \phi^2_{jg}, \xi^2_{jg} &= \pi_{jg}\text{Normal}(0, \phi^2_{jg}) + (1-\pi_{jg})\text{Normal}(0, \xi^2_{jg}), \\ &\text{for } j = 1, \ldots, J \text{ items and } g = 2 \ldots, G \text{ groups.} \end{aligned} \tag{6.5}$$

Here, ϕ^2_{jg} represents the variance of the spike and is fixed to a small number. ξ^2_{jg} is fixed to a large number to represent the variance of the slab. π_{jg} represents the mixing probability for item j. If we set $\pi_{jg} = 1$, the prior consists only of a small-variance normal spike and we thus obtain a similar prior as in approximate MI but now specified for the deviances δ_{jg} instead of the pairwise differences between measurement parameters. Note that the spike-and-slab prior is only specified for the deviances in $G-1$ groups, since this will automatically imply a prior on the deviances in the remaining group. Note that, as a result, the implied marginal prior of the deviances in the remaining group will be more diffuse compared to the other marginal priors.

The mixing probabilities π_{jg} are given a Beta prior, i.e., $\pi_{jg} \sim \text{Beta}(\alpha_j, \beta_j)$. The advantage of the Beta prior lies in the intuitive incorporation of prior information regarding the measurement (non-)invariance. Specifically, the proportion of α_j and β_j reflects the proportion of a priori expected number of invariant and

noninvariant groups. The exact magnitudes of α_j and β_j represent the uncertainty about this expected proportion. For example, if we expect a priori that a measurement parameter is invariant in 20 out of 25 groups, we could simply specify $\alpha_j = 20$ and $\beta_j = 5$ to obtain a Beta prior centered around 0.8. However, we could also specify $\alpha_j = 2$ and $\beta_j = 0.5$ to obtain a Beta prior which still has an expected value equal to 0.8 but is more spread out to reflect the uncertainty about the proportion of invariant and noninvariant groups. Similarly, if we set $\alpha_j = 200$ and $\beta_j = 50$, the Beta prior will be highly peaked around 0.8 and will exert more influence on the posterior. Other priors for the mixing probabilities are possible. For example, George and McCulloch (1993) and Soares et al. (2009) specify $\pi_{jg} \sim \text{Bernoulli}(p_j)$, in which the mixing probabilities take on the value 0 or 1 with some probability p_j.

6.3.5 Robust invariance: the regularized horseshoe prior

The horseshoe prior (Carvalho et al., 2010) is a popular shrinkage prior in the Bayesian literature due to several desirable theoretical properties (see e.g., Carvalho, Polson, & Scott, 2009; Polson & Scott, 2011). Specifically, the horseshoe prior has a global shrinkage parameter that controls general shrinkage of all parameters towards zero, and local shrinkage parameters that control the desirable amount of shrinkage for individual coefficients. This results in heavy shrinkage of relatively small effects and little shrinkage for large effects. In practice, however, the small amount of shrinkage of large effects can be problematic, especially when parameters are weakly identified (Piironen & Vehtari, 2017b). As a result, the posterior means of the parameters might not exist or the MCMC sampler becomes unstable (Ghosh et al., 2018). The regularized horseshoe solves this problem by inducing more shrinkage on large coefficients compared to the horseshoe. It is specified as follows (Piironen & Vehtari, 2017b):[2]

$$\delta_{jg}|\phi_j^2, \tilde{\xi}_{jg}^2 \sim \text{Normal}(0, \phi_j^2 \tilde{\xi}_{jg}^2), \text{ with } \tilde{\xi}_{jg}^2 = \frac{c_{jg}^2 \xi_{jg}^2}{c_{jg}^2 + \phi_j^2 \xi_{jg}^2}, \tag{6.6}$$

$$\begin{aligned} \xi_{jg}|\nu_1 &\sim \text{half-Student's } t_{\nu_1}(0, 1), \\ \phi_j|\nu_2, \tau_j, \sigma_{jg}^2 &\sim \text{half-Student's } t_{\nu_2}(0, \tau_j \sigma_{jg}^2), \\ c_{jg}^2|f, s &\sim \text{inverse-Gamma}(\frac{f}{2}, \frac{fs}{2}) \\ &\text{for } j = 1, \ldots, J \text{ items and } g = 2 \ldots, G \text{ groups.} \end{aligned} \tag{6.7}$$

The resulting prior will shrink small deviances towards zero in a similar way as the original horseshoe. Large deviances, however, will be regularized according to

[2]Note that Piironen and Vehtari (2017b) use different symbols namely: $\phi = \tau$ and $\xi = \lambda$.

a Normal$(0, c^2)$ prior. By specifying the inverse-Gamma prior for c^2, the marginal prior for large deviances becomes a Student's t distribution with f degrees of freedom and scale parameter s.

An additional advantage of the regularized horseshoe is the fact that prior information can be easily incorporated. In the context of a regression model, Piironen and Vehtari (2017b) recommend to choose the scale parameter τ based on a prior guess for the number of relevant predictors p_0, i.e.,

$$\tau = \frac{p_0}{(p - p_0)\sqrt{N}}, \tag{6.8}$$

where p equals the total number of predictors. In the multiple group factor model, the number of predictors translates to the number of groups, with the a priori expected number of relevant predictors being the a priori expected number of groups in which that measurement parameter is assumed to be non-invariant. The sample size N is the total sample size, across all groups [3].

As noted by Piironen and Vehtari (2017b), the original horseshoe prior can be regarded as a continuous version of the spike-and-slab prior with an infinite slab variance. The regularized horseshoe, on the other hand, can be seen as a continuous version of the spike-and-slab prior with a finite slab variance. Contrary to the spike-and-slab prior in (6.5), the regularized horseshoe prior is conditioned on the residual variance of the indicators, σ^2_{jg}.

6.3.6 Comparison of the prior distributions

Figure 6.1 shows the prior densities (left column) and survival functions (right column) corresponding to the different methods for measurement invariance. The densities and survival functions are shown for several settings of the parameters in the prior (i.e., the hyperparameters). The prior densities are useful to determine the behavior around zero, i.e., how much small deviances from the average measurement parameter across groups are shrunken. The more peaked the prior distribution, the more shrinkage towards zero occurs. The survival functions are useful to illustrate the tail behavior. The survival function describes the probability that a parameter has a value greater than the value on the x-axis. For example, at $x = 0$, this probability equals 0.50 for all priors since they are symmetric around zero. The tail behavior determines how robust the method is to large deviations from MI, with heavier tails resulting in more robust methods.

[3] The derivation of τ based on the prior guess in (Piironen & Vehtari, 2017b) is based on a derivation of the implied prior on the effective number of nonzero coefficients in a simple linear regression model.

Exact invariance, in which measurement parameters are restricted to be exactly equal can be viewed as a point mass at zero since it forces all deviances to be exactly zero. It can be seen as the least robust method, not allowing for any deviations from MI. As a result, if there exist violations from MI, exact invariance will lead to biased estimates.

Approximate MI corresponds to a normal distribution which, in the original Mplus specification, is specified for the pairwise differences. As the variance of the normal distribution decreases, the prior becomes more peaked around zero and therefore exerts more shrinkage. In the extreme case, as the variance goes to infinity, the prior becomes uniform on all possible values for the pairwise measurement parameters. This prior does not shrink the estimates whatsoever and will therefore result in a nonidentified model. Thus, the variance needs to be small enough to ensure an identified model. However, as can be seen from the survival function, the smaller the variance, the lighter the tails. Thus, approximate MI shows a tradeoff between model identification and robustness.

The spike-and-slab prior has a spike comparable to approximate MI, in Figure 6.1 a normal spike is shown with a variance of 0.01. Compared to approximate MI, however, the spike-and-slab prior is made more robust by adding a slab component with a larger variance, in this case equal to 1. The spike-and-slab prior in Figure 6.1 is shown for various values of the mixing probability, specifically $\pi_{jg} = c(0.5, 0.9, 0.1)$. As the mixing probability increases, more prior mass will be assigned to the spike and the resulting prior will resemble approximate MI more closely. For small values of the mixing probability, the prior will be more comparable to the slab and, depending on the variance of the slab, might lead to a nonidentified model.

The regularized horseshoe prior has multiple hyperparameters that influence its behavior. Most importantly is the global shrinkage parameter (ϕ_j in (6.6)). As noted by Piironen and Vehtari (2017b), the usual recommendations of $\phi \sim C^+(0, 1)$ or $\phi \sim C^+(0, \sigma^2)$ generally result in values for the global shrinkage parameter that are too large, resulting in a prior that does not shrink the estimates sufficiently. That is why Piironen and Vehtari (2017b) recommend to determine the scale of the prior for the global shrinkage parameter based on prior information in a manner that generally leads to a small scale. In Figure 6.1, the regularized horseshoe prior is plotted for a global scale of 1 and 0.1. A larger global scale leads to a less peaked prior with tails that go to zero more slowly. In addition, the tail behavior is influenced by the hyperparameters of the inverse-Gamma prior on c^2_{jg}. Recall from Subsection 6.3.5 that large deviances will be shrunken according to a Student's t distribution with f degrees of freedom and scale s. Thus, by decreasing the degrees of freedom f or by increasing the scale s, we can obtain a more heavily tailed

regularized horseshoe prior.

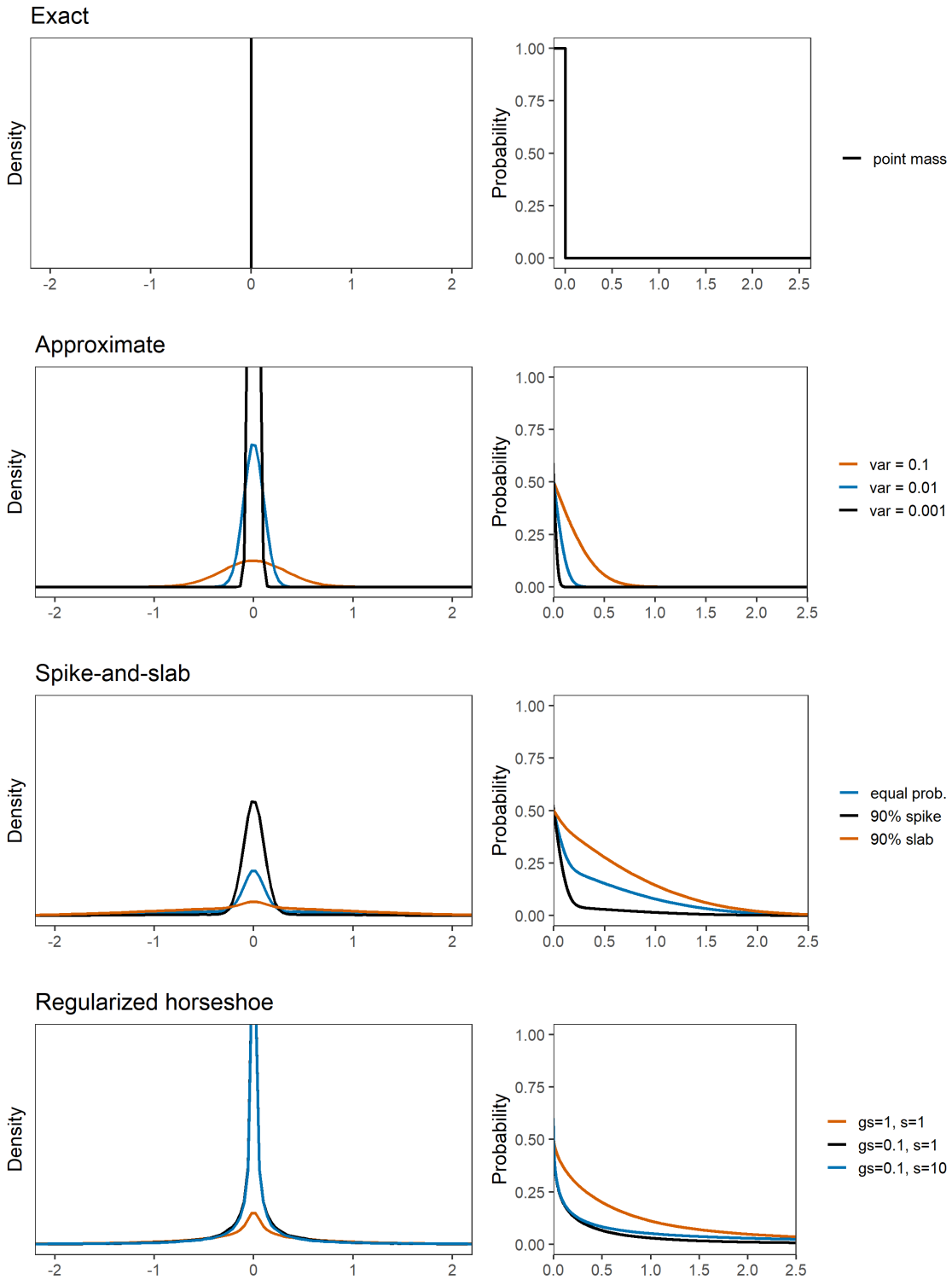

Figure 6.1: Density plots and survival functions of the priors for different hyper-parameter settings. For the regularized horseshoe, "gs" refers to the global scale parameter τ_j, whereas "s" refers to the scale parameter s in the marginal prior for large deviances.

6.4 Illustration

To illustrate the behavior of the shrinkage priors, we analyze different data sets with varying types and degrees of noninvariance. Each data set is simulated using the multiple group factor model in (6.1) with one factor and five items. The factor scores η_{ig} and the measurement errors ϵ_{ijg} of that data set are rescaled to ensure that the parameter values in the data set correspond exactly to the desired parameter values. This way, only one replication for each type of MI is needed. All analyses were run by calling JAGS (Plummer, 2003) from R using `R2jags` (Su & Yajima, 2015). All code is available online at https://osf.io/cp35e/.

6.4.1 Setup

All invariant intercepts are fixed to 0 and all invariant loadings are fixed to 1. The residual variances σ_{jg} are all fixed to 1. We follow a similar setup as in Asparouhov and Muthén (2014) by considering three different types of groups: $\eta_{i1} \sim N(0, 1)$, $\eta_{i2} \sim N(0.3, 2.25)$, and $\eta_{i3} \sim N(1, 1.44)$. The number of groups equals 15, so that there are 5 groups of each type[4]. We consider 5 different conditions for the number of non-invariant items: either 0, 1, 2, 3, or 4 out of 5 items are non-invariant, so that we obtain 0 to 80% of non-invariance. Violations of invariance are induced on the loadings and intercepts simultaneously, for example, in the 20% non-invariance condition the intercept of one item and the loading of one (different) item both violate invariance. We consider approximate violations of invariance only in the 20% non-invariance setting by specifying deviances of the measurement parameters of 0.1. Large violations of MI are considered for all settings. The non-invariant measurement parameters and their values are varied in each group according to five different configurations, which are presented in the Appendix. Since we have 15 groups, these five different configurations are repeated thrice.

6.4.2 Methods

Given that the goal of multiple group factor analysis is generally to compare the factor means and/or variances across groups, we focus on the average absolute error for the factor means and variances. The error is calculated by subtracting the population value from the estimated value (the posterior median) for each group and subsequently averaging the absolute values across groups[5]. We compare the

[4]We have conducted a similar illustration with 60 groups, the results of which are available online at https://osf.io/cp35e/ and did not differ much qualitatively from the results with 15 groups. Only the scalar model with Bayesian estimation performed better with 60 groups and convergence was lower for the shrinkage priors.

[5]Individual estimates and their 95% confidence or credibility intervals are available online at https://osf.io/cp35e/

shrinkage priors to the following methods:

1. Scalar invariance using ML without an anchor item. The latent variable is identified in the first group by fixing the latent mean and variance. The latent variable is linked across groups through exact equality constraints on all intercepts, loadings, and residuals.

2. Scalar invariance using ML with an incorrect anchor item. The latent variable is identified by fixing the factor mean to 0 and the fifth loading in group 1 to 1. The latent variable is linked across groups through exact equality constraints on all intercepts, loadings, and residuals.

3. Scalar invariance using ML with a correct anchor item. The latent variable is identified by fixing the first loading and intercept in group 1. The latent variable is linked across groups through exact equality constraints on all intercepts, loadings, and residuals.

4. Scalar invariance using Bayesian estimation without an anchor item. The latent variable is identified in the first group by fixing the latent mean and variance. The latent variable is linked across groups through exact equality constraints on all intercepts, loadings, and residuals.

5. Bayesian approximate MI. The latent variable is identified in the first group by fixing the latent mean and variance. The latent variable is linked across groups through the prior distribution.

6. Alignment with a Bayesian approximate base model. The latent variable is identified in the first group by fixing the latent mean and variance. The latent variable is linked across groups through the simplicity function.

7. Alignment with a configural base model using ML estimation. The latent variable is identified in the first group by fixing the latent mean and variance. The latent variable is linked across groups through the simplicity function.

For the shrinkage priors, we compare different types of prior information. Specifically, for the spike-and-slab prior, we choose the hyperparameters for the Beta prior on π_{jg} in such a way that the number of non-invariant groups for that parameter is equal to the data-generating values, or that the number of non-invariant groups is equal to half the true number of non-invariant groups. We also consider a setting in which no prior information is included, i.e., $\pi_{jg} \sim \text{Beta}(1, 1)$. We set the precision of the spike equal to 100 and the precision of the slab to 0.01. For the regularized horseshoe, we use similar settings for the scale parameter τ_j, with $\tau_j = 1$ when no prior information is available. The other parameters are chosen such that

we obtain a robust regularized horseshoe specification, specifically: $\nu_1 = 1$, $\nu_2 = 1$, $f = 4$, and $s = 2$.

6.4.3 Convergence

For the shrinkage priors and the scalar invariance model with Bayesian estimation, we concluded that an analysis is converged if the potential scale reduction factor (PSRF) is smaller than 1.2 and the effective sample size is at least 100 for the factor means and variances. For the analyses with Mplus, no effective sample size was available so we based convergence solely on the PSRF, again using a cutoff of 1.2.

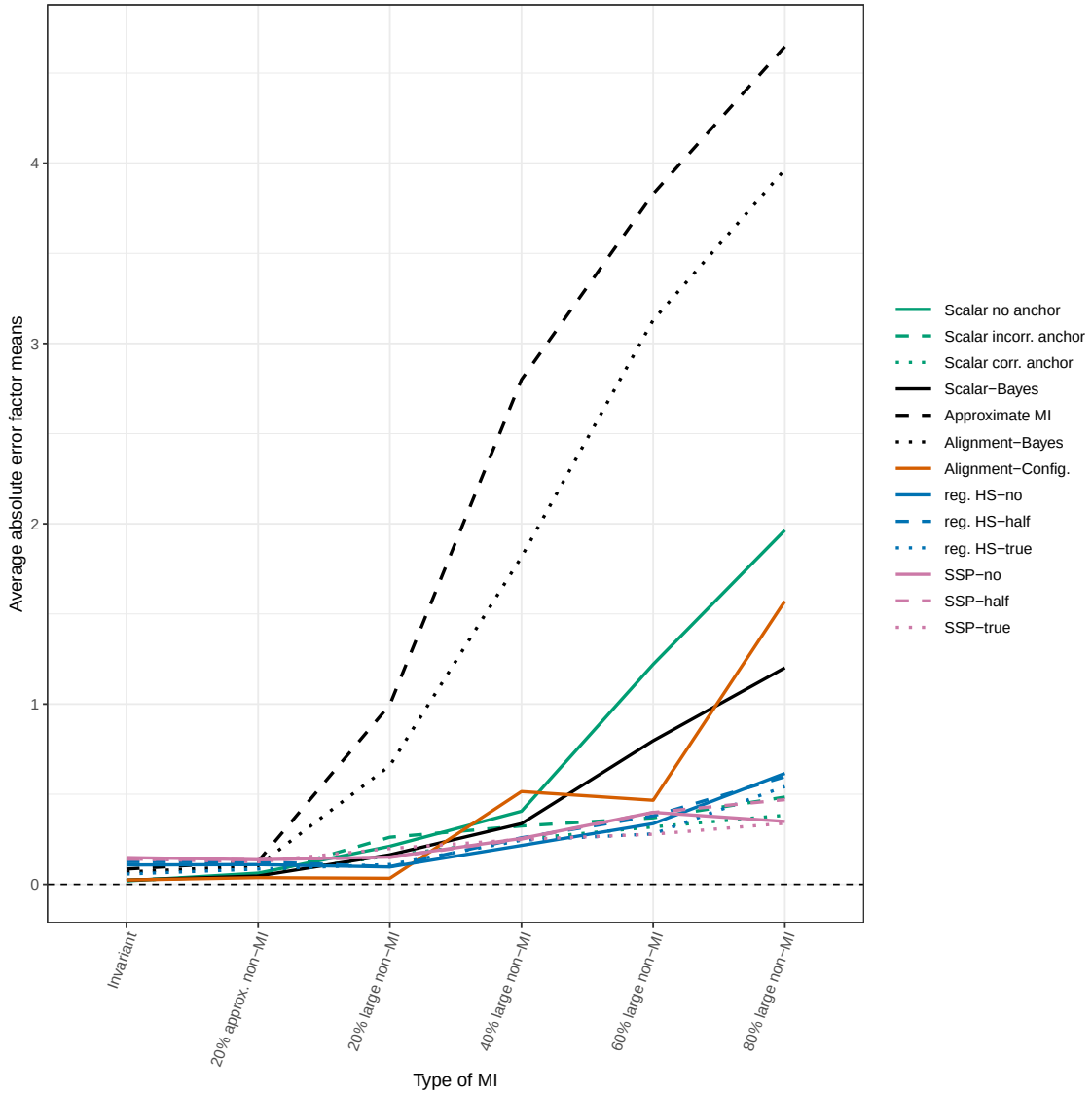

Figure 6.2: Absolute errors (estimate - population value) for the factor means, averaged across the 15 groups

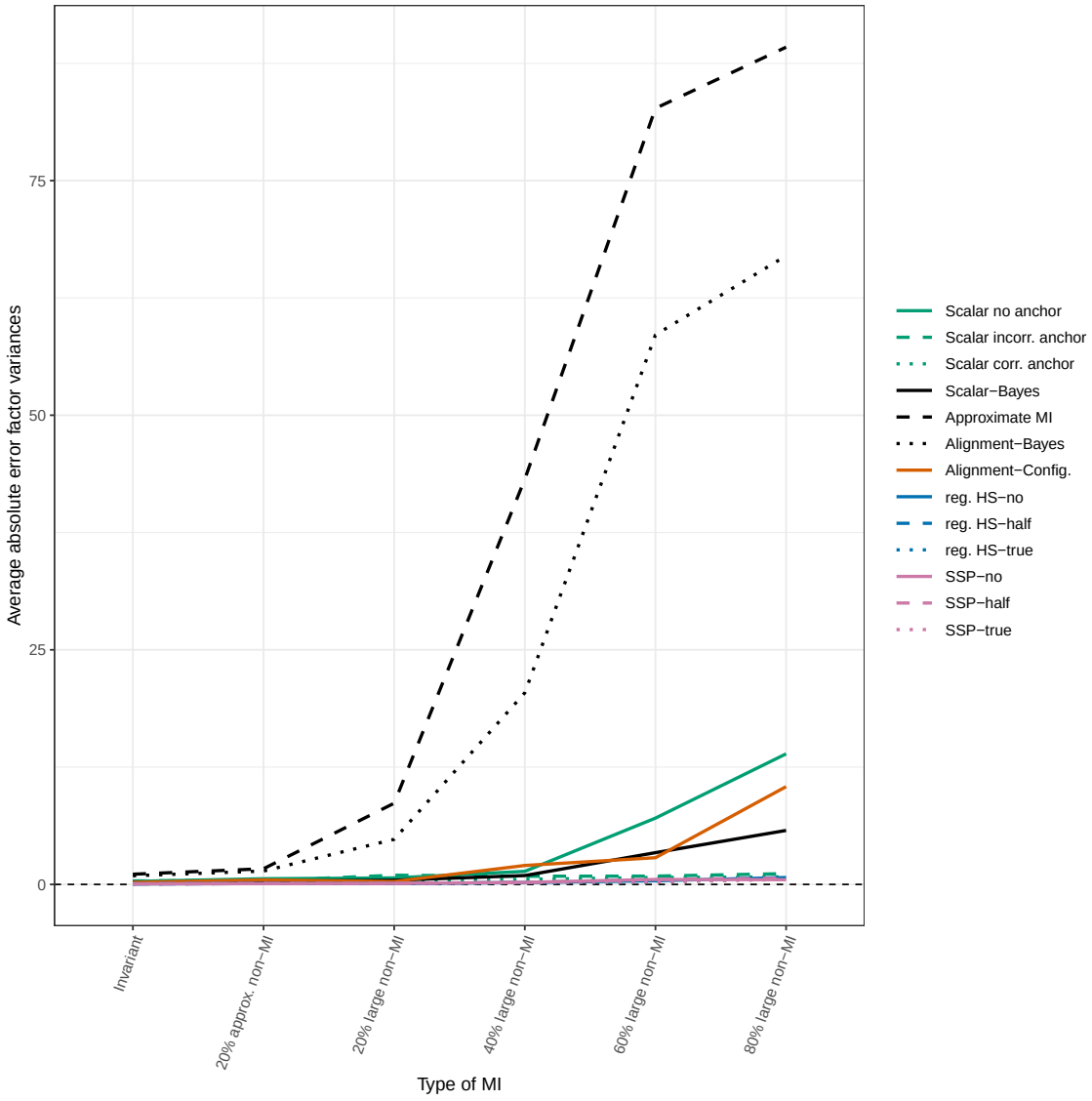

Figure 6.3: Absolute errors (estimate - population value) for the factor variances, averaged across the 15 groups

Figure 6.2 shows the average absolute error in factor means for the different methods, across types of MI when the number of groups equals 15. For the shrinkage priors, two analyses do not reach the convergence criteria when 60% of the items show large violations of MI (SSP-half and SSP-true) and four analyses do not reach the convergence criteria when 80% of the items show large violations (reg. HS-true, SSP-half, SSP-no, SSP-true). In the 60% case, the traceplots for the factor means and variances indicate convergence, with only the factor variance in group 11 showing slight bimodality. In the 80% violations case, multiple posteriors show bimodality. Thus, the results are less trustworthy in these situations. This nonconvergence is not surprising; especially in the 80% large non-MI condition, the violations from MI are extreme. The convergence issues is simply the method indicating that the situation at hand is complex and it might not be wise to compare the factor means.

For the scalar model using Bayesian estimation, all analyses converged. Approximate MI and the alignment method with a Bayesian base model also converged, but since these were analysed with Mplus, this decision was solely based on the PSRF criterium. The traceplots indicate extreme peaks from the 20% large non-MI condition onwards, especially for the factor variances. These peaks become more extreme as the violations get more extreme. This is most likely due to the prior used for the latent variances, which is by default an inverse-Gamma(-1, 0) prior. This is a noninformative improper prior that specifies equal probability mass to all positive values a priori which, in situations with violations from MI, might not be informative enough leading to the unstable MCMC sampler.

6.4.4 Results

For all methods, the error increases when more items are non-invariant. The shrinkage priors show lower average errors than the other methods for large amounts of MI. The differences between the shrinkage priors and their levels of prior information are small. Surprisingly, the scalar model with an anchor item, either correct or incorrect, performs similarly to the shrinkage priors. However, this might be due to the way in which the simulation was set up: for the incorrect anchor item, the factor mean was set to 0 while the fifth loading in group 1 was set to 1. The anchor item is thus incorrect only in terms of the restriction on the loading and not in terms of the restriction on the factor mean. We can expect the average error to be higher when the anchor item is incorrect with respect to both the loading and factor mean restrictions. Moreover, in reality, we do not know the correct anchor item or to what extent the restrictions we impose on the anchor item hold. The scalar models without an anchor item show increased errors from 20% non-invariant items onwards, with ML estimation showing larger errors compared to Bayesian estimation. The alignment method with the configural base model outperforms the shrinkage priors up to 20% of non-invariance but has a higher average absolute error for larger amounts of non-invariance, which is in line with the findings by Asparouhov and Muthén (2014); Flake and McCoach (2018). The alignment method with the Bayesian approximate base model performs much worse than the configural alignment method. This is most likely due to the instability of the MCMC sampler, especially with the default priors used in Mplus. As expected, Bayesian approximate MI only performs well when there is strict or approximate invariance. Figure 6.3 shows the results for the factor variances. The ordering of the methods is similar to that of the factor means, but the errors are generally much larger. This is especially the case for approximate MI and the alignment method with a Bayesian base model, which is most likely due to the default priors used for these analyses.

6.5 Application

To illustrate the use of shrinkage priors to model measurement invariance, we will use data from the fifth wave of the European Social Survey (ESS, 2010). Similar to Asparouhov and Muthén (2014), we focus on four items that measure tradition and conformity, which are combined into one latent variable. The full data set consists of 50,781 observations from 26 countries. To reduce computational time, we focus on a subset of 7 countries. We selected 3 countries with the highest and lowest factor means based on the results in Table 8 from Asparouhov and Muthén (2014). In addition, we included country 22 as reference group. For simplicity, we remove all observations with missing data resulting in 12,811 complete observations. We standardize the items to have a mean of zero and a variance of 1.

We apply both shrinkage priors to this data set. In order to achieve convergence, it was necessary to choose the hyperparameters in such a way that there was enough shrinkage of small deviances of invariance towards zero. [6] Specifically, for the spike-and-slab prior we use a precision for the spike equal to 10,000 and a precision for the slab equal to 1. The mixing probabilities are given the following prior: $\pi_{jg} \sim \text{Beta}(1000, 1)$. For the regularized horseshoe prior, we set the global scale parameter τ_j equal to 0.0001 while the scale of the slab s equals 1. We set all degrees of freedom parameters equal to 3. For the nuisance parameters, we use the same weakly informative priors as specified in Section 6.2, equation 6.3. We compare the results to the alignment method with country 22 as reference group, following (Asparouhov & Muthén, 2014).

6.5.1 Results

First, we consider the posterior mean estimates of the factor means, which are shown in Table 6.1. The ordering of the countries on the latent tradition-conformity variable is the same for both priors and the estimates of the factor means are very similar. Note that the factor means are more pulled towards zero compared to the alignment method. This is most likely due to the quite restrictive hyperparameters needed to achieve convergence. The ordering of the countries is similar between the shrinkage priors and the alignment method, with only the Netherlands and Portugal switched in the alignment method.

[6] Although these hyperparameters were different from the ones used in Section 6.4, we reran the illustration with these hyperparameters and found comparable results, except for small amounts of non-invariance. In those situations, the hyperparameters from the application resulted in lower average absolute errors similar to the alignment method. The results of this sensitivity analysis are available online at https://osf.io/cp35e/.

Table 6.1: Posterior mean estimates factor means with standard deviations in brackets for the application

Country	Alignment (ref. = 22)	Spike-and-slab prior	Regularized horseshoe
23. Sweden (SE)	0.853 (0.081)	0.435 (0.026)	0.438 (0.029)
18. Netherlands (NL)	0.413 (0.050)	0.213 (0.023)	0.219 (0.028)
21. Portugal (PT)	0.437 (0.073)	0.067 (0.021)	0.078 (0.022)
22. Russian Federation (RU)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
24. Slovenia (SI)	-0.297 (0.046)	-0.035 (0.022)	-0.036 (0.021)
13. Greece (GR)	-0.301 (0.042)	-0.148 (0.021)	-0.151 (0.021)
4. Cyprus (CY)	-0.499 (0.051)	-0.184 (0.035)	-0.173 (0.048)

Next, we consider the deviance parameters with a focus on the intercepts, i.e., δ^{ν}_{jg}. Figure 6.4 shows the point estimates and 95% credibility intervals for δ^{ν}_{jg} for both priors as well as the alignment method. The point estimates for the shrinkage priors are based on the posterior mean. Note that the alignment method does not provide estimates for the average deviance from the group mean δ_{jg}, so the estimates in Figure 6.4 were computed manually based on the point estimates for the measurement intercepts by subtracting the average intercept across groups, i.e., $\delta^{\nu}_{jg} = \nu_{jg} - \mu^{\nu}_{j}$. As a result, no confidence interval estimates are available for the alignment method. It is clear that the results do not differ much across the priors, whereas there are several substantial differences between the shrinkage priors and the alignment method. Note that deviances close to zero indicate invariance, whereas values away from zero indicate violations from invariance. In some cases the alignment method indicates invariance while the shrinkage priors indicate a violation from invariance (e.g., the intercept for "imptrad" in Portugal), while in other cases the alignment method indicates non-invariance while the shrinkage priors indicate invariance (e.g., the intercept for "ipfrule" in Portugal). In general, the differences between the shrinkage priors and alignment method are substantial for the intercepts in Portugal and to a lesser extent in Slovenia and Cyprus.

A figure such as Figure 6.4 can be insightful in assessing the amount of non-invariance in each item. Based on Figure 6.4 we can for example conclude, based on the results of the shrinkage priors, that the intercept for the item "ipfrule" is invariant in all countries except the Netherlands (NL) and Greece (GR). A similar figure for the loading parameters is available in Figure 6.5 and shows, for example, that the loading for the item "ipfrule" is invariant in all countries. In general, the differences between the shrinkage priors and alignment method are much smaller for the loadings compared to the intercepts.

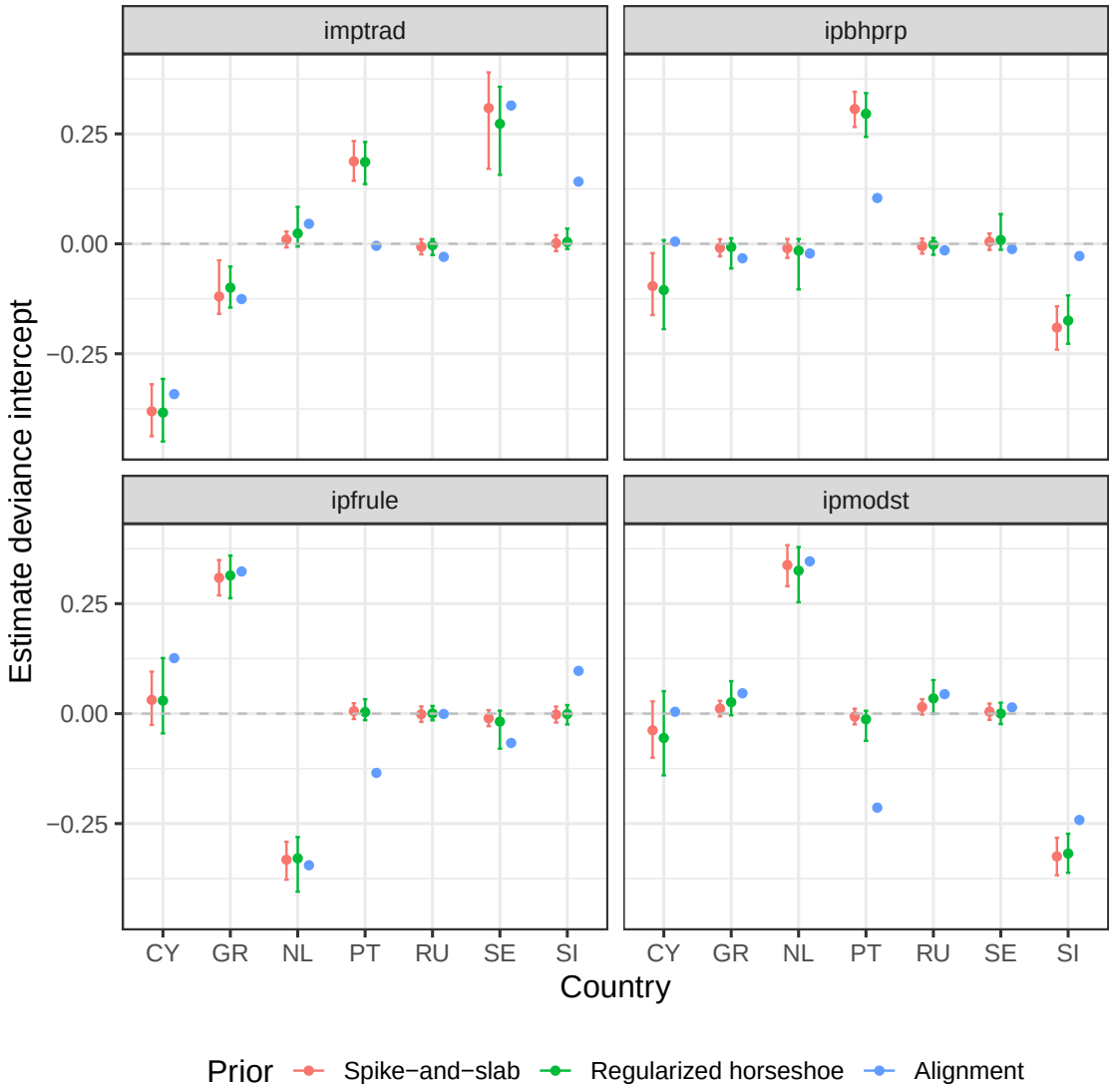

Figure 6.4: Point estimates and 95% credibility intervals for the average deviances of the intercepts in the application

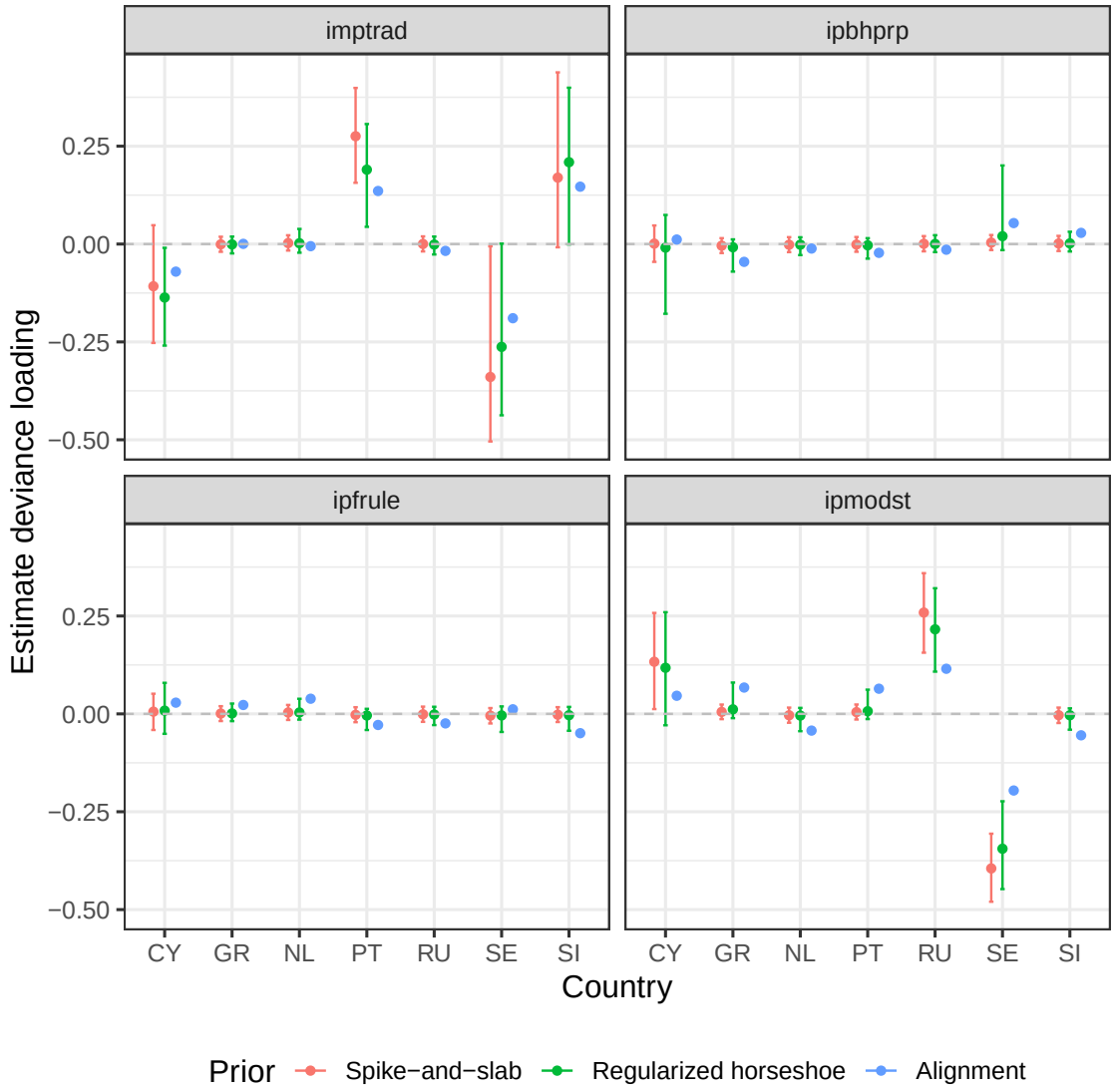

Figure 6.5: Point estimates and 95% credibility intervals for the average deviances of the loadings in the application

6.5.2 Sensitivity of the results

With any Bayesian analysis, it is of importance to assess the sensitivity of the results to the chosen prior distributions. Although the results show no sensitivity to the specific distributional form used (i.e., spike-and-slab or regularized horseshoe prior), the analysis is sensitive to the specific choice of the hyperparameters. Specifically, for the spike-and-slab prior we tried a less restrictive version of the prior in which the mixing probabilities were given the hyperprior $\pi_{jg} \sim \text{Beta}(500, 1)$, $\pi_{jg} \sim \text{Beta}(100, 1)$, or $\pi_{jg} \sim \text{Beta}(10, 1)$. However, these analyses resulted in convergence issues, specifically bimodal posteriors for various parameters including the factor means. This bimodality persisted even when the precision of the spike was further increased to 100,000. For the regularized horseshoe, increasing the global

scale parameter τ_j to 0.001 for the loadings (based on the fact that the loadings showed more invariance compared to the intercepts in Asparouhov and Muthén (2014)) also resulted in bimodal posteriors as did decreasing the degrees of freedom to 1. A second factor that can influence the results is the chosen reference group. In order to identify the factor, we need to fix the factor mean in one country to 0 and the average loading for one item to 1. Following Asparouhov and Muthén (2014), we chose country 22 as reference group. However, we tried to run the analysis with group 23 as reference group, but this analysis did not converge for both priors. Note that for the "fixed" alignment method, the factor means are also sensitive to the choice of the reference group. This is not the case for the "free" alignment method, however, as mentioned by (Asparouhov & Muthén, 2014) this approach cannot always be used due to identification issues.

Overall, we can conclude from the application that while the shrinkage priors provide very similar results in terms of factor means and invariance of the measurement parameters, these results differ substantially from those obtained using the alignment method. This is especially the case for the intercepts and less so for the factor loadings. Overall, the intercepts also show more violations of invariance: based on the shrinkage priors, 10 out of 28 intercepts are noninvariant. This is around 36% which is more than the 25% of violations that is assumed by the alignment method. Based on the results of the illustration in Section 6.4, we might put more trust in the results of the shrinkage priors compared to those of the alignment method. Still, the shrinkage priors also show an increased error in the estimation of the factor means as the amount of noninvariance increases. In practice, it is therefore recommended to conduct a simulation study based on the data set at hand to investigate how reliable estimates are for the parameters of interest (such as the factor means) given the data-specific characteristics.

6.6 Discussion

In this paper we presented a novel method to model measurement invariance based on the use of robust shrinkage priors. Specifically, we focused on the regularized horseshoe and spike-and-slab prior. The robust shrinkage priors automatically identify the multiple group factor model, without requiring unrealistic restrictions. We illustrated the behavior of the priors and showed how they can improve the estimates of the factor means and variances compared to the alignment method when large amounts of noninvariance are present. In the empirical application we showed how the shrinkage priors can be applied in practice and how the posteriors of the deviance parameters might be used to assess which measurement parameters show violations of invariance. In the empirical application, the estimates of the factor

means differed between the alignment methods and the shrinkage priors. This is not surprising given the differences between the approaches. First of all, the alignment method uses a pairwise parametrization while the shrinkage priors are specified on each deviance from the average value across groups. Moreover, quite restrictive hyperparameters were needed for the shrinkage priors to achieve convergence in the empirical application. Future research should compare the methods further in less extreme situations. In addition, it might be of interest to specify the shrinkage priors on the pairwise differences in measurement parameters to provide a closer comparison to the alignment method. It would be interesting to find out what type of prior distribution corresponds exactly to the alignment method and compare this prior distribution to the robust shrinkage priors.

Although the shrinkage priors offer a very promising method for measurement invariance, there are still several limitations to overcome. First of all, the method can be quite slow especially as the number of groups increases. In addition, the results can become unstable for larger number of groups. We experienced this when trying to apply the shrinkage priors to the full ESS data of 26 countries. Running this analysis with the same hyperparameters as used in Section 6.5 for 25,000 iterations took approximately 13.7h for the spike-and-slab prior and 4.4h for the regularized horseshoe. However, 25,000 iterations was not enough to achieve convergence, with several posteriors being bimodal including some of the factor means. This is most likely due to the fact that the hyperparameters are not restrictive enough for that specific data set, i.e., there are too many violations of invariance so restrictive hyperparameters are needed to identify the model. However, given the large computation time, it becomes impractical to tune the hyperparameters in such large data sets. To solve this problem, it might be worthwhile to investigate the possibility of parallelizing the computations further. Currently, only the chains are run in parallel but further parallelizing the computations within each chain could provide a significant speedup. This can be done, for example, using the `map_rect` function in Stan (Carpenter et al., 2017). A second limitation is the sensitivity of the results to the reference group used to identify the factor. This might be solved by fixing the average intercept across groups for one item to 0 instead of one factor mean. A further improvement of the restrictions of the model would be to replace the sum-to-zero constraint on the deviances for each item in the first group, since the marginal prior of the restricted parameter will be more diffuse compared to the other marginal priors and the choice of the parameter to restrict is arbitrary. See Rouder, Morey, Speckman, and Province (2012) for an alternative approach.

Throughout this paper, we have focused on the multiple-group confirmatory factor model, which views the groups as a fixed source of variation and limits the inference to the groups in the sample. Recently, a random approach relying on mul-

tilevel models has gained popularity, especially due to its ability to include variables that might explain why violations of invariance arise (Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012). In the random approach, the inference is to the population from which the groups were drawn. In certain situations, the random approach is preferable over the fixed approach (B. O. Muthén & Asparouhov, 2017). It would therefore be interesting to generalize the shrinkage priors to be applicable within the random multilevel approach to measurement invariance as well.

In general, the use of shrinkage priors to model measurement invariance offers exciting new possibilities. Future research should further investigate the strengths and limitations of this approach using more extensive simulation studies than the one considered in this paper. Moreover, more formal methods for assessing the violations of invariance can be developed, for example based on the posterior of the mixing probabilities in the spike-and-slab prior. Finally, we have focused solely on the spike-and-slab and the regularized horseshoe priors, but many other shrinkage priors exist that can be applied in the context of measurement invariance.

Appendix: Set-up illustration

Values for the measurement parameters in the illustration

Group	Non-invariant intercepts	Values	Non-invariant loadings	Values
		20% approximate violations		
1	ν_1	-0.1	λ_2	0.9
2	ν_2	0.1	λ_3	1.1
3	ν_3	-0.1	λ_4	0.9
4	ν_4	0.1	λ_5	1.1
5	ν_5	-0.1	λ_5	0.9
		20% large violations		
1	ν_1	-0.5	λ_2	0.3
2	ν_2	-0.3	λ_3	0.5
3	ν_3	-0.1	λ_4	0.7
4	ν_4	0.3	λ_5	1.3
5	ν_5	0.5	λ_5	1.5
		40% large violations		
1	ν_1, ν_2	-0.5, -0.5	λ_2, λ_3	0.3, 0.3
2	ν_2, ν_3	-0.3, -0.3	λ_3, λ_4	0.5, 0.5
3	ν_3, ν_4	-0.1, -0.1	λ_4, λ_5	0.7, 0.7
4	ν_4, ν_5	0.3, 0.3	λ_5, λ_2	1.3, 1.3
5	ν_5, ν_1	0.5, 0.5	λ_5, λ_3	1.5, 1.5
		60% large violations		
1	ν_1, ν_2, ν_3	-0.5, -0.5, -0.5	λ_2, λ_3, λ_4	0.3, 0.3, 0.3
2	ν_2, ν_3, ν_4	-0.3, -0.3, -0.3	λ_3, λ_4, λ_5	0.5, 0.5, 0.5
3	ν_3, ν_4, ν_5	-0.1, -0.1, -0.1	λ_4, λ_5, λ_2	0.7, 0.7, 0.7
4	ν_4, ν_5, ν_1	0.3, 0.3, 0.3	λ_5, λ_2, λ_3	1.3, 1.3, 1.3
5	ν_5, ν_1, ν_2	0.5, 0.5, 0.5	λ_5, λ_3, λ_4	1.5, 1.5, 1.5
		80% large violations		
1	ν_1, ν_2, ν_3, ν_4	-0.5, -0.5, -0.5, -0.5	λ_2, λ_3, λ_4, λ_5	0.3, 0.3, 0.3, 0.3
2	ν_2, ν_3, ν_4, ν_5	-0.3, -0.3, -0.3, -0.3	λ_3, λ_4, λ_5, λ_2	0.5, 0.5, 0.5, 0.5
3	ν_3, ν_4, ν_5, ν_1	-0.1, -0.1, -0.1, -0.1	λ_4, λ_5, λ_2, λ_3	0.7, 0.7, 0.7, 0.7
4	ν_4, ν_5, ν_1, ν_2	0.3, 0.3, 0.3, 0.3	λ_5, λ_2, λ_3, λ_4	1.3, 1.3, 1.3, 1.3
5	ν_5, ν_1, ν_2, ν_3	0.5, 0.5, 0.5, 0.5	λ_5, λ_3, λ_4, λ_2	1.5, 1.5, 1.5, 1.5

Chapter 7

Epilogue

Bayesian structural equation modeling is increasingly popular due to the advantages it offers over classical, maximum likelihood based SEM. Multiple studies have compared Bayesian and classical SEM and concluded that BSEM performs better in terms of convergence (Kohli et al., 2015), inadmissible estimates (Can et al., 2014; Dagne et al., 2002), bias (Depaoli & Clifton, 2015), requires smaller sample sizes (Hox et al., 2012), and can handle more complex models (Harring et al., 2012; Lüdtke et al., 2013; Oravecz et al., 2011). Additionally, the flexibility offered by BSEM further extends the usefulness of the SEM framework. For example, BSEM allows straightforward credible intervals not only on parameters themselves but also on functions of them (see e.g., Geldhof et al., 2014; Y. Yuan & MacKinnon, 2009). Moreover, through incorporation of prior information, traditional restrictions such as cross-loadings or error covariances can be relaxed leading to more realistic models (B. O. Muthén & Asparouhov, 2012). The aim of this thesis was to investigate whether this belief in BSEM is warranted. We did so by investigating the most important component of any Bayesian analysis: the prior distribution.

Often, researchers wish to use the capabilities of BSEM without including any substantive information in the analysis. As a result, most applications of BSEM rely on "default" priors; objective priors that can be used in an automatic fashion. In Chapter 2, we investigated various such default priors in the context of SEM. We focused on priors for the variance parameters, since those are highly sensitive to the prior. We considered various noninformative improper priors, vague proper priors, and empirical Bayes priors, with especially the first two categories often being used, for example as default prior setting in software such as Mplus. The results indicate that, especially for small samples, not one default prior performs best and Bayesian estimation with default priors does not outperform ML estimation. This is not surprising since the default priors do not include any information over-and-above the information in the data, which is the information used by ML estimation.

To see if we might improve the performance of BSEM, Chapter 3 looks into more robust priors for random effects variance parameters in a multilevel structural equation model. By including prior information in the analysis, we aim to improve the performance of BSEM compared to the default priors considered in Chapter 2. However, in practice, prior information might not be available or might be incorrectly incorporated into the prior distribution. Therefore, we compared weakly informative, correctly specified informative, and incorrectly specified informative priors in Chapter 3. Overall, the priors considered in Chapter 3 did not appear sufficiently informative to counteract the small number of groups used in the simulation studies. The parameter of interest in MLSEMs, the contextual effect, showed bias across the board and the power to detect a small contextual effect was much too low, even with 50 groups and informative priors. Moreover, the informative pri-

ors resulted in convergence issues, making a thorough comparison difficult. Overall, Chapter 3 can be seen as a starting point for future investigations into robust informative prior distributions for BSEM.

Another way in which the prior distribution might be employed to improve the performance of the model is by specifying the prior in such a way that it automatically imposes the restrictions needed to identify the model. By doing so, more realistic models can be specified. Shrinkage priors are especially suited for this purpose, since they have the ability to automatically shrink small effects towards zero, thereby imposing the needed restrictions on the model, while simultaneously keeping large effects large. However, since there exist many different shrinkage priors, it is unclear which shrinkage priors would be best suited to apply in the context of BSEM. Therefore, in Chapters 4 and 5 we have investigated the available shrinkage priors in the context of simple linear regression models. Chapter 4 presented a theoretical overview of a selection of commonly used shrinkage priors and showed that most shrinkage priors perform similarly when the number of predictors is smaller than the sample size. When the number of predictors exceeded the sample size, some evidence suggests that the more sophisticated global-local shrinkage priors such as the (regularized) horseshoe and hyperlasso perform best. Chapter 5 translated the findings from Chapter 4 into a practical guide on how to use and tune the various priors in regression models when the sample size is small, with a specific focus on software.

Taken together, Chapters 4 and 5 provided the basis for Chapter 6 in which two shrinkage priors, the spike-and-slab prior and the regularized horseshoe, were applied to model measurement invariance. The multiple group factor model generally used to model MI is a prime example of a model in which restrictions are needed to identify the model. Traditionally, these restrictions are imposed and, if necessary, freed in an ad-hoc manner. Shrinkage priors, on the other hand, can automatically incorporate the required restrictions in the model during estimation. Chapter 6 investigated this property for the spike-and-slab prior and the regularized horseshoe in a multiple group factor model and found that these priors offer promising new ways to model MI. Nevertheless, more research is still needed to further increase the practical usefulness of these methods.

Overall, we can conclude from this thesis that BSEM has great potential, which arises mainly from the prior distribution. However, the results presented in this thesis also warn against naive use of the prior distribution, for example by relying blindly on noninformative default specifications, which can lead to worse performance compared to classical estimation methods (see also, Smid et al., 2019). Therefore, the main recommendation of this thesis is to take care when specifying the prior distribution in BSEM. To do so, first of all ensure that you understand

how your prior behaves before you conduct the Bayesian analysis. The chapters in this thesis aim to help this understanding. Plotting the prior distribution with various hyperparameter settings can offer insight into how to tune the prior for each parameter. Moreover, prior predictive checks in which simulations from the prior predictive distribution are visualized are important to understand the behavior of the joint prior over all parameters simultaneously (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019). Second, after you have done the analysis with the specified prior distribution, always perform a sensitivity analysis to assess the sensitivity of the results to the specific prior setting that has been used. The goal of this prior sensitivity analysis is to further your understanding of the prior distribution and to understand whether the prior distribution was indeed specified as intended. In Chapter 2, we have provided guidelines on how to conduct such a sensitivity analysis for default BSEM. There, we have relied on rerunning the analysis with different prior distributions, however, there exist automatic ways to assess prior sensitivity as well (see, for example, Gustafson & Wasserman, 1995; Roos, Martins, Held, & Rue, 2015). Future research should investigate such automatic approaches in the context of BSEM.

With a thorough understanding of the prior distribution, it can be used to advance the field of SEM. An example is provided in Chapter 6 in which we use the characteristics of the shrinkage priors to obtain a more flexible and realistic multiple group factor model. This idea of using regularization or penalization in SEM is already quite popular in classical SEM (Jacobucci, Brandmaier, & Kievit, 2019; Jacobucci, Grimm, & McArdle, 2016) and gaining popularity in BSEM (Feng, Wang, et al., 2017; Feng, Wu, & Song, 2017; Jacobucci & Grimm, 2018; Lu et al., 2016). Bayesian regularized SEM is one of the most exciting areas for future research. The main challenge is to make the approach more practical in terms of tuning, compation time, and selecting which parameters should be restricted to zero. How to tune the shrinkage priors depends on the model and data at hand, and should result in a prior that restricts the model enough to identify it while not creating bias due to the restriction of parameters that should in reality be free. This process is complicated by the unstability of the MCMC sampler as the model becomes non-identified as well as the large computation time generally needed. Unlike classical penalization approaches such as the lasso, Bayesian regularization approaches do not automatically restrict parameters to be exactly zero. Thus, an additional step is needed in Bayesian regularization to restrict small parameters to zero. In Chapters 4 and 5 we relied on marginal credible intervals to do so, but it would be better to consider the joint posterior distribution (Piironen et al., 2017). Future research should adapt approaches for joint variable selection to the context of BSEM (for example, projection predictive variable selection; Piironen & Vehtari,

2017a, or decoupled shrinkage and selection; Hahn & Carvalho, 2015).

Other areas for future research include the incorporation of survey weights into BSEM as well as investigations into missing data handling. Although it is possible to model the inclusing probabilities to solve the first issue (see e.g., Little, 2015) and to automatically impute the missing data within the estimation procedure (see e.g., Gelman et al., 2013, Chapter 18), these model extensions further complicate the model which can make estimation problematic (Gelman, 2007). Moreover, available software programs for BSEM are currently not yet capable to handle the modeling of inclusion probabilities, missing data, as well as more sophisticated prior distributions such as shrinkage priors. Thus, despite its popularity, the field of BSEM remains in its infancy and can still be greatly improved to become more reliable, robust, and practical.

References

Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian adaptive lasso quantile regression. *Statistical Modelling*, *12*(3), 279–297. doi: 10.1177/1471082x1101200304
111, 122

Andersen, M. R., Vehtari, A., Winther, O., & Hansen, L. K. (2017). Bayesian inference for spatio-temporal spike-and-slab priors. *Journal of Machine Learning Research*, *18*(139), 1-58.
111

Arbuckle, J. L. (2013). *IBM SPSS Amos 22 user's guide.*
16

Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica.* doi: 10.5705/ss.2011.048
112

Asparouhov, T., & Muthén, B. O. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology* (pp. 2531–2535).
7, 62

Asparouhov, T., & Muthén, B. O. (2010). *Bayesian analysis of latent variable models using Mplus.* Retrieved from www.statmodel.com/download/BayesAdvantages18.pdf
7

Asparouhov, T., & Muthén, B. O. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Retrieved from https://www.statmodel.com/download/NCME12.pdf
62

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. doi: 10.1080/10705511.2014.919210
169, 170, 174, 180, 184, 185, 189

Asparouhov, T., & Muthén, B. O. (2019). Latent variable centering of predic-

tors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. doi: 10.1080/10705511.2018.1511375
7, 67

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using big data to understand the human condition: The kavli HUMAN project. *Big Data*, *3*(3), 173–188. doi: 10.1089/big.2015.0012
110

Bae, K., & Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, *20*(18), 3423–3430. doi: 10.1093/bioinformatics/bth419
112

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151-164. doi: 10.1037/a0030642
14

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58-80. doi: 10.1214/088342304000000116
17, 27

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *3*, 385-402. doi: 10.1214/06-ba115
9, 15, 20, 21, 23, 48, 57, 63, 111, 169

Berger, J. O., & Strawderman, W. E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *The Annals of Statistics*, *24*(3), 931-951. doi: 10.1214/aos/1032526950
21, 63

Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle*. Hayward, CA: Institute of Mathematical Statistics.
56

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*. Retrieved from https://arxiv.org/abs/1701.02434
134

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017a, dec). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105–1131. doi: 10.1214/16-ba1028
112

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. T. (2017b). Lasso meets horseshoe: A survey. *arXiv preprint arXiv:1706.10179*. Retrieved from

https://arxiv.org/abs/1706.10179
112

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2012). Bayesian shrinkage. *arXiv preprint arXiv:1212.6088*. Retrieved from https://arxiv.org/abs/1212.6088
112

Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, *45*(3), 370. doi: 10.2307/2095172
6, 19

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.
19

Bornn, L., Gottardo, R., & Doucet, A. (2010). Grouping priors and the Bayesian elastic net. *arXiv preprint arXiv:1001.4083*. Retrieved from https://arxiv.org/abs/1001.4083
112

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*, 173-187. doi: 10.1007/s11222-013-9424-2
133

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. doi: 10.1214/06-ba117
82

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi: 10.1037/0033-2909.105.3.456
169

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01
149, 155

Can, S., van de Schoot, R., & Hox, J. (2014). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement*, *75*(3), 406-427. doi: 10.1177/0013164414547959
10, 14, 195

Carlin, B. P., & Louis, T. A. (2000a). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). New York: Chapman and Hall/CRC.

10, 15, 20, 23, 24, 25, 56

Carlin, B. P., & Louis, T. A. (2000b). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, *95*(452), 1286-1289. doi: 10.1080/01621459.2000.10474331

24

Caron, F., & Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on machine learning - ICML '08*. Association for Computing Machinery (ACM). doi: 10.1145/1390156.1390168

112

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

16, 154, 190

Carter, G., & Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, *69*(348), 880-885. doi: 10.1080/01621459.1974.10480222

24

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Aistats* (Vol. 5, pp. 73–80). Retrieved from http://www.jmlr.org/proceedings/papers/v5/carvalho09a/carvalho09a.pdf

176

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. doi: 10.1093/biomet/asq017

112, 117, 123, 124, 156, 176

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83-87. doi: 10.2307/2682801

15, 20, 24

Casella, G. (1992). Illustrating empirical Bayes methods. *Chemometrics and Intelligent Laboratory Systems*, *16*(2), 107-125. doi: 10.1016/0169-7439(92)80050-e

24

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071-4089. doi: 10.1002/sim.5821

57

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *In A. Brito and J. Teixeira Eds., Proceedings of the 5th FUture BUsiness TEChnology Conference* (pp. 5–12).

140

Dagne, G. A., Howe, G. W., Brown, C. H., & Muthén, B. O. (2002). Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. *Psychological Methods*, *7*(2), 262. doi: 10.1037/1082-989x.7.2.262
10, 14, 195

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, *27*(3), 567-578. doi: 10.2307/3316112
106

Darnieder, W. F. (2011). *Bayesian methods for data-dependent priors* (Unpublished doctoral dissertation). The Ohio State University.
25, 56

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. doi: 10.1177/0022022112438397
191

Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(2), 178–203. doi: 10.1080/10705511.2012.659614
9, 14, 15

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological methods*, *18*(2), 186-219. doi: 10.1037/a0031609
9, 14, 15

Depaoli, S. (2014). The impact of inaccurate informative priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 239-252. doi: 10.1080/10705511.2014.882686
9, 14, 15

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-25. doi: 10.1080/10705511.2014.937849
9, 10, 14, 15, 64, 77, 104, 195

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. doi: 10.1037/met0000065
7, 15, 45, 46

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise

variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. doi: 10.1111/j.2044-8317.1992.tb00992.x
110

Dua, D., & Graff, C. (2019). *UCI machine learning repository.* Retrieved from http://archive.ics.uci.edu/ml
155

Dunson, D. B., Palomo, J., & Bollen, K. A. (2005). *Bayesian structural equation modeling* (Tech. Rep.). SAMSI. Retrieved from http://www.samsi.info/sites/default/files/tr2005-05.pdf
18, 24, 49

Efron, B. (1996). Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association*, *91*(434), 538-550. doi: 10.1080/01621459.1996.10476919
25, 56

Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika*, *82*(2), 295–307. doi: 10.1007/s11336-017-9564-y
173

ESS. (2010). *ESS round 5: European social survey round 5 data.* Data file edition 3.3. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC.
185

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2017). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. doi: 10.3758/s13423-017-1317-5
7

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273
120

Fawcett, T. (2015). Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data*, *3*(4), 249–266. doi: 10.1089/big.2015.0049
110

Feng, X.-N., Wang, Y., Lu, B., & Song, X.-Y. (2017). Bayesian regularized quantile structural equation models. *Journal of Multivariate Analysis*, *154*, 234–248. doi: 10.1016/j.jmva.2016.11.002
112, 149, 170, 197

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2015). Bayesian adaptive lasso for ordinal regression with latent variables. *Sociological Methods & Research*. doi: 10.1177/0049124115610349
122, 133

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 341–358. doi: 10.1080/10705511.2016.1257353
197

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, *1*(1), 1–40. doi: 10.1214/06-ba101
7

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 56–70. doi: 10.1080/10705511.2017.1374187
170, 184

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi: 10.18637/jss.v033.i01
133

Frühwirth-Schnatter, S., & Wagner, H. (2010). Stochastic model specification search for gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, *154*(1), 85–100. doi: 10.1016/j.jeconom.2009.07.003
68

Fúquene, J., Pérez, M.-E., & Pericchi, L. R. (2014). An alternative to the inverted Gamma for the variances to modelling outliers and structural breaks in dynamic models. *Brazilian Journal of Probability and Statistics*, *28*(2), 288–299. doi: 10.1214/12-bjps207
69

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402. doi: 10.1111/rssa.12378
197

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701. doi: 10.1198/016214505000000105
9, 14

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*,

19(1), 72-91. doi: 10.1037/a0032138
14, 195

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515-534. doi: 10.1214/06-ba117a
15, 21, 23, 28, 31, 57, 63, 68, 69, 77, 114

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164. doi: 10.1214/088342306000000691
198

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.
169, 198

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman and Hall/CRC.
7, 27

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360-1383. doi: 10.1214/08-aoas191
57

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733-807.
47

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472. doi: 10.1214/ss/1177011136
47, 83, 134

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881. doi: 10.2307/2290777
117, 126, 156, 175, 176

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383. doi: 10.1214/17-ba1051
69, 125, 176

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231. doi: 10.1207/s15328031us0104_02
67

Griffin, J. E., & Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, *12*(1), 135–159. doi: 10.1214/15-ba990

112

Griffin, J. E., & Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *Working paper. University of Warwick. Centre for Research in Statistical Methodology*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.3430&rep=rep1&type=pdf

117, 119, 156

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, *53*(4), 423–442. doi: 10.1111/j.1467-842x.2011.00641.x

117, 122, 123, 128, 156

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*(1), 124–136. doi: 10.1037/0022-0663.95.1.124

64

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Lüdtke, O., & Diallo, T. M. O. (2019). A systematic evaluation and comparison between exploratory structural equation modeling and Bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–28. doi: 10.1080/10705511.2018.1554999

174

Gustafson, P. (2010). Bayesian inference for partially identified models. *The International Journal of Biostatistics*, *6*(2). doi: 10.2202/1557-4679.1206

10

Gustafson, P., Hossain, S., & Macnab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, *34*(3), 377–390. doi: 10.1002/cjs.5550340302

106

Gustafson, P., & Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, *23*(6), 2153-2167. doi: 10.1214/aos/1034713652

197

Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, *110*(509), 435–448. doi: 10.1080/01621459.2014.993077

149, 198

Hallquist, M., & Wiley, J. (2014). MplusAutomation: Automating Mplus model estimation and interpretation [Computer software manual]. Retrieved from

http://CRAN.R-project.org/package=MplusAutomation (R package version 0.6-3)
31, 46

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845. doi: 10.1093/biomet/asp047
111, 112

Harring, J. R., Weiss, B. A., & Hsu, J.-C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological methods*, *17*(2), 193-214. doi: 10.1037/a0027539
14, 195

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press.
111, 115, 152, 155

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. doi: 10.3102/0162373707299706
77

Heerwegh, D. (2014). Small sample Bayesian factor analysis. In *Phuse*. Retrieved from http://www.lexjansen.com/phuse/2014/sp/SP03.pdf
27

Helm, C. (2018). How many classes are needed to assess effects of instructional quality? A Monte Carlo simulation of the performance of frequentist and Bayesian multilevel latent contextual models. *Psychological Test and Assessment Modeling*, *60*(2), 265–285.
64, 77, 104

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*(436), 1461-1473. doi: 10.1080/01621459.1996.10476714
21

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.2000.10485983
111, 118

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research*, *26*(3), 329–367. doi: 10.1177/0049124198026003003
6

Hox, J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equa-*

tion Modeling: A Multidisciplinary Journal, *8*(2), 157–174. doi: 10.1207/s15328007sem0802_1
6, 62, 82

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*(2), 87-93.
10, 14, 27, 63, 79, 104, 195

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *The Statistician*, *24*(4), 267. doi: 10.2307/2987923
117, 118, 156

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. doi: 10.1214/009053604000001147
112, 126, 175

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science*, *2*(1), 55–76. doi: 10.1177/2515245919826527
197

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–11. doi: 10.1080/10705511.2017.1410822
149, 170, 197

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. doi: 10.1080/10705511.2016.1154793
197

Jak, S. (2018). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–16. doi: 10.1080/10705511.2018.1534205
67

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. doi: 10.1007/bf02289343
5

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.
14

Kaseva, T. (2018). *Convergence diagnosis and comparison of shrinkage priors*. Retrieved from `https://livefull.github.io` (Github repository)
132, 135

Kass, R. E., & Natarajan, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 535-542. doi: 10.1214/06-ba117b
24

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343-1370. doi: 10.1080/01621459.1996.10477003
20

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley, 1979. (Revised edition available from: http://davidakenny.net/books.htm)
5

Klein, N., & Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, *11*(4), 1071–1106. doi: 10.1214/15-ba983
63

Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear–linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological methods*, *20*(2), 259-275. doi: 10.1037/met0000034
10, 14, 195

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. doi: 10.1214/10-ba607
111, 117, 122, 132, 141, 154

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963-974. doi: 10.2307/2529876
24

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? a simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*(15), 2401–2428. doi: 10.1002/sim.2112
15

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.
48

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons, Ltd.
7

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*(4), 653-686. doi: 10.1207/s15327906mbr3904_4
27

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. doi: 10.1214/10-ba506
111, 117, 121, 132, 133, 154, 156

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. doi: 10.1198/016214507000001337
24

Lichman, M. (2013). *UCI machine learning repository.* Retrieved from http://archive.ics.uci.edu/ml
140, 143

Little, R. J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the IAOS*, *31*(4), 555–563. doi: 10.3233/sji-150944
198

Liu, H., Xu, X., & Li, J. J. (2017). HDCI: High dimensional confidence interval based on lasso and bootstrap [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=HDCI (R package version 1.0-2)
141

Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*(4), 519–539. doi: 10.1080/00273171.2016.1168279
112, 149, 170, 175, 197

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological methods*, *16*(4), 444-467. doi: 10.1037/a0024376
14

Lüdtke, O., Robitzsch, A., Kenny, D. A., & Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychological methods*, *18*(1), 101-119. doi: 10.1037/a0029252
14, 195

Lumley, T. (2017). leaps: Regression subset selection [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=leaps (R package version 3.0)

133

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. doi: 10.1002/sim.3680

63

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

16, 20

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. doi: 10.1027/1614-2241.1.3.85

14

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, *17*(3), 340-345. doi: 10.1037/a0027131

14

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. doi: 10.1037/0033-2909.111.3.490

169

Makalic, E., & Schmidt, D. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649*. Retrieved from https://arxiv.org/abs/1611.06649

155

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, *1*(2), 133–163. doi: 10.1111/j.1745-6916.2006.00010.x

64

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*(6), 764–802. doi: 10.1080/00273170903333665

62, 65, 77

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*(3), 319–350. doi: 10.1007/s10648-008-9075-6

64

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. doi: 10.1016/0005-2795(75)90109-9

134

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. doi: 10.1080/00273171.2015.1036965

110, 152

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

137

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4). doi: 10.18637/jss.v085.i04

16, 19

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*(1), 45-58.

6, 14, 62

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

117, 119, 156

Miočević, M., Levy, R., & Savord, A. (2020). The role of exchangeability in sequential updating of findings from small sample studies. In M. van de Schoot R. & Miočević (Ed.), *Small sample size solutions: A guide for applied researchers and practitioners* (chap. 2).

153

Miočević, M., Levy, R., & van de Schoot, R. (2020). Introduction to Bayesian statistics. In M. van de Schoot R. & Miočević (Ed.), *Small sample size solutions: A guide for applied researchers and practitioners* (chap. 1).

153

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. doi: 10.1080/01621459.1988.10478694

117, 126, 156, 175

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2016). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in*

Ecology and Evolution, *8*(3), 339–348. doi: 10.1111/2041-210x.12681
134

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/sim.8086
83, 86

Mulder, J., & Fox, J.-P. (2013). Bayesian tests on components of the compound symmetry covariance matrix. *Statistics and Computing*, *23*, 109-122. doi: 10.1007/s11222-011-9295-3
56

Mulder, J., & Fox, J.-P. (2018). Bayes factor testing of multiple intraclass correlations. *Bayesian Analysis*. doi: 10.1214/18-ba1115
106

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887-906. doi: 10.1016/j.jspi.2009.09.022
25, 26

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530-546. doi: 10.1016/j.jmp.2009.09.003
25

Mulder, J., & Pericchi, L. R. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, *13*(4), 1189–1210. doi: 10.1214/17-ba1092
69, 114

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. doi: 10.1177/0049124194022003006
62, 67

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi: 10.1037/a0026802
6, 7, 14, 44, 195

Muthén, B. O., & Asparouhov, T. (2013). *BSEM measurement invariance analysis*. Retrieved from http://statmodel.com/examples/webnotes/webnote17.pdf
169, 174

Muthén, B. O., & Asparouhov, T. (2017). Recent methods for the study of mea-

surement invariance with many groups. *Sociological Methods & Research*, 004912411770148. doi: 10.1177/0049124117701488
191

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide. seventh edition.* Los Angeles, CA: Muthén and Muthén.
16, 20, 21, 47

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599-620. doi: 10.1207/s15328007sem0904_8
32

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: the big-fish-little-pond-effect for PISA 2006. *Educational Psychology*, *31*(5), 629–656. doi: 10.1080/01443410.2011.586416
64

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*(449), 227-237. doi: 10.1080/01621459.2000.10473916
15, 20, 24, 26, 106

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45-60. doi: 10.1093/pan/mpt014
47

Oberski, D. L. (2019). Rank-deficiencies in a reduced information latent variable model. *arXiv preprint arXiv:1911.00770*. Retrieved from https://arxiv.org/abs/1911.00770
5

OECD. (2007). *Pisa 2006*. doi: 10.1787/9789264040014-en
64

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological methods*, *16*(4), 468-490. doi: 10.1037/a0024375
14, 195

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. doi: 10.1198/016214508000000337
113, 117, 119, 153, 156, 158

Peltola, T., Havulinna, A. S., Salomaa, V., & Vehtari, A. (2014). Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on*

Bayesian Modeling Applications Workshop-Volume 1218 (pp. 79–88).
112

Pérez, M.-E., Pericchi, L. R., & Ramírez, I. C. (2017). The scaled beta2 distribution as a robust prior for scales. *Bayesian Analysis*, *12*(3), 615–637. doi: 10.1214/16-ba1015
69

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *163*(7), 670–675. doi: 10.1093/aje/kwj063
134

Piironen, J., Betancourt, M., Simpson, D., & Vehtari, A. (2017). Contributed comment on article by van der Pas, Szabó, and van der Vaart. *Bayesian Analysis*, *12*(4), 1264-1266.
149, 164, 197

Piironen, J., & Vehtari, A. (2015). Projection predictive variable selection using Stan+R. *arXiv preprint arXiv:1508.02502*. Retrieved from https://arxiv.org/abs/1508.02502
69, 164

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. doi: 10.1007/s11222-016-9649-y
149, 197

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. doi: 10.1214/17-ejs1337si
125, 156, 171, 176, 177, 178

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124).
16, 180

Polson, N. G., & Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian statistics 9* (pp. 501–538). Oxford University Press (OUP). doi: 10.1093/acprof:oso/9780199694587.003.0017
112, 120, 124, 176

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. doi: 10.1214/12-ba730
57, 63, 68, 69, 114, 124

Polson, N. G., Scott, J. G., & Windle, J. (2014). The Bayesian bridge. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), *76*(4), 713–733. doi: 10.1111/rssb.12042
120

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/
19

Raftery, A. E., & Lewis, S. (1991). *How many iterations in the Gibbs sampler*. Retrieved from http://people.ee.duke.edu/~lcarin/raftery92how.pdf
47

Redmond, M., & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, *141*(3), 660–678. doi: 10.1016/s0377-2217(01)00264-8
143, 152, 155

Revilla, M., & Saris, W. E. (2013). The split-ballot multitrait-multimethod approach: Implementation and problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 27–46. doi: 10.1080/10705511.2013.742379
5

Rindskopf, D. (1984). Structural equation models. *Sociological Methods & Research*, *13*(1), 109–119. doi: 10.1177/0049124184013001004
5

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementations*. New York, NY: Springer Science+Business Media.
48, 56

Roos, M., Held, L., et al. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, *6*(2), 259–278. doi: 10.1214/11-ba609
68

Roos, M., Martins, T. G., Held, L., & Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, *10*(2), 321-349. doi: 10.1214/14-ba909
197

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. doi: 10.18637/jss.v048.i02
19

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5),

356–374. doi: 10.1016/j.jmp.2012.08.001
190

Roy, V., & Chakraborty, S. (2016). Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis*. doi: 10.1214/16-ba1025
112, 114, 132

Rubin, D. B. (1980). Empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, *75*(372), 801-816. doi: 10.1080/01621459.1980.10477553
24

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583–601. doi: 10.1080/10705510903203466
14

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561-582. doi: 10.1080/10705510903203433
47

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52. doi: 10.1007/bf02294318
7

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28. doi: 10.1214/16-sts576
63, 70

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–31. doi: 10.1080/10705511.2019.1577140
106, 196

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348–377. doi: 10.3102/1076998609332752
171, 175, 176

Stan Development Team. (2015). *Prior choice recommendations.* https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. (Accessed: 2019-08-05)
71

Stan Development Team. (2016). *rstanarm: Bayesian applied regression modeling via Stan.* (R package version 2.13.1)
149, 155

Stan Development Team. (2017a). Stan Modeling Language Users Guide and Reference Manual, *version 2.17.0.* Retrieved from http://mc-stan.org
134

Stan Development Team. (2017b). The Stan Core Library, *version 2.16.0.*
133

Stan Development Team. (2018). *RStan: the R interface to Stan.* (R package version 2.18.2)
83, 133

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107. doi: 10.1086/209528
168

Su, Y.-S., & Yajima, M. (2015). *R2jags: Using R to run JAGS.* (R package version 0.5-7)
180

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
110, 111, 120, 132

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. doi: 10.1111/j.1467-9868.2011.00771.x
111

Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 148–159. doi: 10.1111/j.2517-6161.1974.tb00996.x
9, 14

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3). doi: 10.18637/jss.v045.i03
156

van der Pas, S., Szabó, B., & van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, *12*(4), 1221–1274. doi: 10.1214/17-ba1065
164

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. O. (2013). Facing off with Scylla and Charybdis: A comparison of scalar,

partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00770
169

van de Schoot, R., Veen, D., Smeets, L., Winter, S., & Depaoli, S. (2020). A tutorial on using the WAMBS-checklist to avoid the misuse of Bayesian statistics. In M. van de Schoot R. & Miočević (Ed.), *Small sample size solutions: A guide for applied researchers and practitioners* (chap. 3).
153, 154

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*, *4*(21). doi: 10.1037/met0000100
9, 15, 16

van de Wiel, M. A., Beest, D. E. t., & Münch, M. (2017). Learning from a lot: Empirical bayes in high-dimensional prediction settings. *arXiv preprint arXiv:1709.04192*. Retrieved from https://arxiv.org/abs/1709.04192
114, 115

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. doi: 10.1037/met0000162
63, 113, 172

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. doi: 10.1016/j.jmp.2018.12.004
161, 162, 170

Veen, D., & Egberts, M. (2020). The importance of collaboration in Bayesian analyses with small samples. In M. van de Schoot R. & Miočević (Ed.), *Small sample size solutions: A guide for applied researchers and practitioners* (chap. 4).
153, 154

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.* (R package version 2.0.0)
115

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 646–648. doi: 10.1093/biomet/74.3.646
116

Wolpert, D. H., & Strauss, C. E. M. (1996). What Bayes has to say about the evidence procedure. In *Maximum entropy and Bayesian methods* (pp. 61–78). Springer Netherlands. doi: 10.1007/978-94-015-8729-7_3
114

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with

grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. doi: 10.1111/j.1467-9868.2005.00532.x
121

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. doi: 10.1037/a0016972
9, 14, 62, 195

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (p. 233-243). Amsterdam, The Netherlands: Elsevier.
24

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (2013). Bayesian inference and application of robust growth curve models using student's t distribution. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 47–78. doi: 10.1080/10705511.2013.742382
57

Zhao, S., Gao, C., Mukherjee, S., & Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, *17*(196), 1-47.
112

Zhou, X., & Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician*, *64*(2), 159–163. doi: 10.1198/tast.2010.09109
65

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 661–679. doi: 10.1080/10705511.2016.1207179
63, 64, 77

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi: 10.1198/016214506000000735
120, 122

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x
111, 121, 132

Summary

The aim of this thesis was to investigate the use of prior distributions in Bayesian structural equation modeling. To this end, various prior distributions have been investigated in the context of different models.

We start our investigation by focusing on default priors in a structural equation model with a mediating effect between the latent variables. Default priors are classes of priors that do not contain any external substantive information, are completely dominated by the information in the data, and can be used in an automatic fashion for a Bayesian data analysis. Multiple default priors exist, which can be roughly divided into three categories: 1) non-informative improper priors; 2) vague proper priors; and 3) empirical Bayes priors. In Chapter 2, various priors from each of these categories have been investigated in a simulation study. Despite their popularity, the results show that different default priors lead to varying results, especially for small sample sizes. This leads to perhaps the most important recommendation in this thesis: to always conduct a prior sensitivity analysis. By doing so, the researcher can assess to what extent the results of the analysis are influenced by the specific (default) prior that is used. In Chapter 2, we provide guidelines for conducting such a prior sensitivity analysis for a default BSEM analysis.

Given the sensitivity of the results to the various default priors, the investigation is continued with more robust prior distributions. Robust priors are characterized by distributional properties, such as heavier tails, that make them less susceptible to unduly influencing the results. In Chapter 3, various robust prior distributions are investigated in the context of a multilevel SEM. The focus lies on the random effects variance parameters, since it is well known that this type of parameter is especially sensitive to the choice of the prior. The amount of information entailed in the robust priors is varied to obtain three different specifications: 1) a default specification that can be used without adaptation in any application; 2) an informative specification that is in line with the population values; and 3) an informative specification that deviates from the population values. Although extensive comparisons between the various robust prior specifications were complicated by convergence issues, the differences between the robust priors were generally small. Instead, differences in the results of the simulation studies are mostly due to the

population conditions. For example, when the population values for the random effects variances equal 1, the uniform and robust priors all show substantial negative bias and coverage rates that are much too low. This indicates that there are limits to the robustness offered by the heavy tails of the robust prior distributions.

The second part of this thesis focuses not on how to diminish the influence of the prior on the results, but rather on how to utilize the prior distribution in a beneficial way. Specifically, the focus lies on so-called shrinkage priors, which are a class of priors that shrink small effects towards zero, while simultaneously keeping large effects away from zero. As a result, shrinkage priors can automatically perform variable selection, which makes them popular in regression models with many predictors. Generally, shrinkage priors are heavily peaked around zero (to shrink small effects) and have heavy tails (to keep substantial effects large). However, many different shrinkage priors exist that fit this general description. Chapter 4 provides an overview of the most popular shrinkage priors in the literature, namely the ridge, local Student's t, lasso, elastic net, group lasso, hyperlasso, (regularized) horseshoe, and discrete normal mixture priors. These priors are compared in terms of prediction and variable selection accuracy within a linear regression context. The results indicate only small differences between the priors and classical penalization approaches when the number of predictors is smaller than the number of observations. Only in the condition in which the number of predictors exceeded the number of observations did the differences between the methods become more pronounced.

In order to improve the usability of the shrinkage priors in practical applications, Chapter 5 translates the findings of Chapter 4 in a manner that is more accessible for applied researchers. It explains the usefulness of penalization methods in general and Bayesian penalization methods in particular and provides an overview of the shrinkage priors with a specific focus on how to set the hyperparameters of these priors as well as how to choose between the various shrinkage priors. Furthermore, Chapter 5 discusses several software packages that can be used for Bayesian penalized regression: `rstanarm`, `brms`, and `bayesreg`, and illustrates the latter package through an applied example.

Although Chapters 4 and 5 examine the shrinkage priors in linear regression models, the ultimate goal is to apply the shrinkage priors in structural equation models. A first step towards this goal is taken in Chapter 6 in which the spike-and-slab prior and the regularized horseshoe prior are used in a multiple group confirmatory factor model. By doing so, a more robust method is obtained to model measurement invariance. Specifically, the shrinkage priors are specified for the deviances from the average value over the groups for each intercept and loading parameter. We compare this novel method to exact and approximate measurement invariance and to the alignment method. An illustration shows how the shrinkage priors can

improve the estimates of the factor means and variances compared to approximate measurement invariance and the alignment method when large amounts of noninvariance are present. Application of the shrinkage priors to data from the European Social Survey illustrates how the posteriors of the deviance parameters might be used to assess which measurement parameters indicate violations of invariance.

Overall, the results of this thesis can be broadly divided into the following three conclusions:

1. Given the sensitivity of the results of a Bayesian SEM analysis to the choice of the (default) prior, a prior sensitivity analysis should always be conducted.

2. The use of heavy-tailed robust priors is recommended over certain default priors to avoid sensitivity of the results to the prior, however, robust priors still require careful specification of the hyperparameters.

3. The use of shrinkage priors in Bayesian SEM offers a promising area for future research.

Acknowledgements

Writing a PhD-thesis is no small feat and I count myself lucky to have had so many people support me throughout this process. I would like to take this opportunity to thank at least a portion of the people without whom this thesis would not exist.

First of all, I am extremely grateful to the Netherlands Organisation for Scientific Research (NWO) for funding this project. It is such a privilege to have your own research ideas funded, allowing you to work on a project you are extremely passionate about.

Importantly, I would never have been awarded the Research Talent Grant if it were not for the support from many people in the department who gave feedback on the proposal and helped me prepare for the interview. The project itself would never have existed, nor would it be successfully completed if not for my (co-)promotors. Jeroen, your expertise in writing grant proposals was indispensable while we were applying for the Research Talent Grant. In addition, I have you to thank for my knowledge regarding categorical data analysis, which you taught me in a Research Master course (despite me being the only student). Also, as head of the department, you played an important role in creating a pleasant and supportive work environment.

I am very lucky to have had not one, but two amazing co-promotors. With Joris being the Bayesian expert and Daniel being the SEM expert, you complemented each other perfectly. Daniel, you were most involved in the beginning of the project whilst you were still working at Tilburg University. In addition to learning a lot from you about SEM, I have you to thank for many general research skills, such as setting up a simulation study, keeping a literature database, and creating insightful figures. This project was not an easy one, but your ability to reframe difficult concepts and put them into a different, easier to understand perspective really helped me in my learning process. Once you started working at Utrecht University, we saw less of each other although you were always an email away if I needed any help.

Joris, you introduced me to Bayesian statistics whilst I was still in the Research Master. Your enthusiasm drew me to this (to me) unknown area of statistics

and your supervision of my First Year Paper and Master Thesis helped me gain a basic understanding of Bayesian statistics, which only further increased my fascination with the subject. During these projects I already noticed what an excellent supervisor you are and this was proven again throughout my PhD. You provided me with a lot of guidance at the beginning of the project but, importantly, let me find my own way as the project progressed, to allow me to become an independent researcher. Throughout the years you were always very quick in providing me with detailed and constructive feedback, making sure to point out the positive points as well, which really helped me in remaining confident about my abilities. I truly could not have wished for a better daily supervisor!

Chapter 3 of this thesis was written whilst I was visiting the Centre for Multilevel Modelling at Bristol University. I am very grateful to everyone there who made me feel welcome. Bill, I very much enjoyed our collaboration. Thank you for all the discussions about the project that would generally end up at some different topic entirely, but also for inviting me to your home for a Sunday roast and a walk through the English countryside.

I would like to thank the reading committee, Prof. Dr. Yves Rosseel, Prof. Dr. Rens van de Schoot, Prof. Dr. Maurits Kaptein, Dr. Rogier Kievit, and Dr. Suzanne Jak, for taking time out of your busy schedules to read, evaluate, and discuss this thesis.

Working on this thesis would not have been half as much fun if it were not for my colleagues. Academia can be quite competitive, which makes the supportive and friendly atmosphere at the department extra special. Whenever I got stuck on a problem, I would only have to wait until 12 pm sharp when we would get together for lunch, during which there would be no talk about work but a lot of laughter instead. After these lunch breaks, I would get back to my desk and find that I could easily solve the problem I had run into before.

In particular, I would like to thank the secretaries, Marieke and Anne-Marie, for all your help and support. Guy, Jelte, and Kim, I would like to thank you for your trust in me to teach Applied Methods and Statistics. I could not have done this without the support of Guy, Wilco, Leonie, and Esther. I would also like to thank all the members of the pubquiz team "Parameters Bier", with whom I had several very entertaining Tuesday nights which inevitably ended in our team being last. Finally, I would like to thank my roommates: Eva, Laura, Davide, and later Andrea, Diana, and Marlyne. Thank you for all the talks, fun, and gossip!

When you are passionate about your work, it is very easy to get lost in it. My family has therefore been crucial in reminding me that there is more to life than my thesis. I am very fortunate that, for the past 12 years, I could count on the support of not one, but two families. Peter Sr., Corry, Eva, and Stan, thank you for letting

me be a part of your family and for all the dinners, drinks, and family weekends that allowed me to relax and recharge.

Mom and dad, thank you for your never-ending and unconditional love and support. From an early age, you instilled in me the importance of doing something that you love, that you are passionate about. Thanks to the freedom you gave me in finding my own path, I have found it. I hope I will be able to raise Nora the same way.

Frank, Joyce, and Hans, no matter how stressed I would be, you were always able to make me laugh, often at my own expense. Marieke, you inspired me to pursue a career in academia in the first place. I am profoundly grateful to you and Paul for helping me every step of the way, whether it was just by providing me with general tips or information or by giving me detailed feedback on applications.

Last, but not least, I want to thank the person without whom I could not have written this thesis. Dear Peter, in the 12 years that we have been together, you have always supported me. However, in the 4,5 years of my PhD, I have required a bit more support from you than usual since writing a thesis can be a stressful process at times. Fortunately, you were always there for me to help me get out of my head and take a break. Thank you for all your patience with me, for always cheering me up and making me laugh, for being interested in my work (particularly in the "Gucci-prior"), and for making me immensely happy! I used to think that I could not get any happier or luckier, but then Nora was born. Now, together we can embark on the adventure of raising her. I am looking forward to it!